FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Estudo e caraterização dos hábitos de utilização e navegação em jornais online

João Miguel Falcão Morgado

DISSERTAÇÃO

Mestrado Integrado em Engenharia Informática e Computação

Orientador: Sérgio Sobral Nunes

21 de Julho de 2017

Estudo e caraterização dos hábitos de utilização e navegação em jornais online

João Miguel Falcão Morgado

Mestrado Integrado em Engenharia Informática e Computação

21 de Julho de 2017

Resumo

Olhando para o estado em que a Internet se encontra nos dias de hoje, em que tudo é desenvolvido com vista a uma experiência de navegação o mais intuitiva possível e até com conteúdos adaptados aos gostos ou necessidades dos utilizadores, torna-se fundamental compreender o modo como os utilizadores interagem com a informação que lhes é apresentada. Nesta dissertação pretende-se estudar e caracterizar os hábitos de utilização e navegação em jornais online, tendo como caso de estudo o JPN, analisando a interação dos utilizadores com conteúdos noticiosos ao longo de um período superior a 10 anos, praticamente desde a fundação do JPN até aos dias de hoje.

Foi feita investigação na área do *Web Mining*, especialmente *Web Usage Mining*, sendo depois identificadas as métricas fundamentais para caracterizar os hábitos de utilização e navegação num jornal online e, posteriormente, recolhidos os dados correspondentes no Google Analytics. O grande intervalo temporal é de extrema importância para perceber como os hábitos e os próprios utilizadores mudaram ao longo do tempo. De forma paralela a esta recolha de dados, foi efetuado um inquérito a estudantes de jornalismo para compreender quais as métricas valorizadas por quem está na área do jornalismo.

Analisando os dados recolhidos é possível verificar que o jornal é consultado maioritariamente na região do Porto, por pessoas entre os 25 e os 34 anos. Consegue-se observar, também, como evoluiram as redes sociais e os dispositivos móveis, com a percentagem de acessos às páginas do JPN a aumentar 23% a partir de redes sociais e 45% a partir de dispositivos móveis, desde 2006. Os horários de consulta também mudaram com o passar dos anos. Em 2006 a hora com mais movimento era entre 9h e 10h - com 5,9% das visitas diárias, diminuindo a partir daí até atingir 1,7% às 23h. Em 2016 o pico é atingido às 15h, com 6,4% das visitas diárias, diminuindo muito pouco até às 23h, quando as visitas ocorridas são 5,75% do total diário. Também o tempo médio de cada sessão aumentou, passando de 73 segundos, em 2005, para 158 segundos, em 2017, o que corresponde a um aumento de cerca de 46%. O maior número de conteúdos multimédia é o responsável pelas sessões com maior duração. Este estudo vem confirmar as tendências que se observam no dia-a-dia, como é o caso da cada vez maior utilização dos dispositivos móveis. Perceber a forma como os utilizadores interagem com a informação é fundamental para os jornalistas se adaptarem à realidade que os rodeia e assim manter os seus leitores fiéis ao jornal. Com a maior utilização dos dispositivos móveis é recomendável que as páginas continuem a ser pensadas para uma fácil leitura neste tipo de dispositivos. Este cuidado com as páginas vai desde o seu *layout* até à orientação das imagens - há que ter em conta que uma imagem na vertical tem uma visualização mais fácil num dispositivo móvel.

Abstract

Looking at the state of the Internet nowadays, where everything is developed with the goal of providing a navigation experience as intuitive as possible and even with content that suits the likes and needs of users it becomes a fundamental matter to understand the way users interact with the information presented to them. In this dissertation, it is intended to study and characterize the habits of use and navigation throughout online newspapers, having as a case study the JPN, analysing the interaction of users with news content over a period of more than 10 years, virtually since the establishment of JPN to this day.

An investigation was carried out around Web Mining, especially Web Usage Mining, after which the fundamental metrics to characterize the habits of use and navigation throughout online newspapers were identified and the corresponding data was collected using Google Analytics. The large time interval is extremely important to understand how the habits and the users themselves have changed over the times. In parallel with the data gathering an inquire was made to journalism students to understand which metrics are valued by those who are in this area.

Analysing the data collected it is possible to realize that the newspaper is mainly consulted in the Porto region by people between 25 and 34 years old. It is also possible to observe how social networks and mobile devices have evolved with a percentage of accesses to the JPN pages increasing 23% from social networks and 45% from mobile devices, since 2006. The schedules of consultation also have changed over the years. In 2006, the busiest hour was between 9am and 10am with 5,9% of daily views, decreasing until 1,7% at 11pm. In 2016 the peak is reached at 3pm, with 6,4% of daily views, decreasing very little until 11pm when the registered visits are 5,75% of the daily total. Also, the average time for each session has increased going from 73 seconds in 2005 to 158 seconds in 2017, an increase of 46%. The largest number of multimedia content is responsible for this increasing. This study confirms an everyday life tendency which consists of the increasing use of mobile devices. Understanding the way users interact with information is paramount for journalists to adapt the reality that surrounds them and thus keeping their readers loyal to the newspaper.

Agradecimentos

Para a elaboração deste projeto tenho de agradecer a certas pessoas, que me motivaram, ajudaram ou trocaram impressões e que, com a sua ajuda, tornaram a sua elaboração muito mais fácil.

Em primeiro lugar, ao meu orientador, o Professor Sérgio Nunes, que se mostrou sempre disponível para o pronto esclarecimento de qualquer dúvida e que procurou ajudar-me a atingir um melhor trabalho através de sugestões.

Ao Pedro Candeias, técnico multimédia do JPN, que me ajudou nos primeiros passos no Google Analtics.

À professora Isabel Reis, diretora do JPN, e à Filipa Silva, editora no JPN, que me ajudaram na mudança de perspetiva de engenharia para jornalismo e que permitiu identificar aspetos importantes para os jornalistas e editores.

À minha família, especialmente aos meus pais e irmã que acompanharam o desenvolvimento de todo o projeto e procuraram motivar-me para as etapas que se iriam seguir.

Aos meus amigos, que se mostraram interessados no desenvolvimento do trabalho, e contribuiram com sugestões.

João Morgado

"Once we accept our limits, we go beyond them"

Albert Einstein

Conteúdo

Lista de Figuras

Lista de Tabelas

Abreviaturas e Símbolos

HTTP	Hypertext Transfer Protocol
IP	Internet Protocol
JPN	Jornalismo Porto Net
URL	Uniform Resource Locator
WUM	Web Usage Mining
SO	Sistema Operativo
OOXML	Office Open XML
JPN	JornalismoPortoNet
CSV	Comma-Separated Values

Capítulo 1

Introdução

Hoje em dia a Internet encontra-se num estado em que tudo é feito de forma a captar o maior número de utilizadores possível. Para tal, é preciso conhecer os seus gostos e hábitos para que, não só a informação que lhes chega seja do seu agrado, como também terem uma experiência de navegação intuitiva. Tão ou mais importante do que captar utilizadores, é mantê-los. Para isso é necessário conhecer a evolução do mundo que nos rodeia e estar consciente dos gostos dos utilizadores não é suficiente para a sua manutenção. É necessário perceber de que forma os utilizadores acompanham essa evolução, para que também seja possível acompanhar as suas necessidades e ter uma estrutura e conteúdos o mais atualizados possível.

Este trabalho insere-se na área do *Web Usage Mining*, uma das três categorias de *Web Mining* [JKA13], que procura extrair conhecimento de um conjunto de dados semi-estruturados que contém as características dos utilizadores/informação [JK98]. É feita uma análise aos utilizadores do JornalismoPortoNet (JPN) [JPNa] (um jornal multimédia de atualização permanente que é, também, um projeto da Licenciatura em Ciências da Comunicação da Universidade do Porto), através da investigação de mais de 10 anos de registos de acessos.

1.1 Motivação

Conhecer o público do jornal *online* é particularmente importante em grandes redações, onde existe uma enorme pressão para vender as suas notícias e onde a concorrência é bastante feroz. De onde vêm as visitas à página do jornal? Quem as faz? A que horas são feitas? Estas são algumas das perguntas a que é importante responder de forma a perceber quem são os destinatários do jornal para que seja possível a sua modelação, tanto a nível estrutural como de conteúdos, de acordo com o tipo de leitores.

Tudo isto são preocupações diárias para quem está na área do jornalismo, mas não se pode deixar o passado cair no esquecimento, uma vez que é ele quem mostra a evolução dos leitores. A intuição leva-nos a pensar que, nos dias que correm, é necessário direcionar os conteúdos para

serem vistos em dispositivos móveis - mudando a orientação em que as fotografias são tiradas, por exemplo - uma vez que estes têm cada vez mais uso no dia a dia. Mas será que as visitas a partir destes dispositivos têm aumentado assim tanto, ao ponto de operar esta mudança? Será que se tem mantido estável nos últimos anos e a melhor opção será a continuidade?

O estudo dos hábitos de navegação em jornais *online* tem benefícios para diferentes públicos: leitores e responsáveis pelo jornal. No caso dos leitores é benéfico uma vez que têm uma experiência de navegação mais intuitiva e personalizada e apenas vão ver aquilo que de facto querem, ou podem querer. Já para os responsáveis, tendo conhecimento e prevendo os hábitos de navegação de quem visita a página, passa a ser possível saber, por exemplo, promover/aumentar o número de leitores e garantir a fidelização desses leitores.

1.2 Objetivos

O objetivo do trabalho é caracterizar os hábitos de navegação dos utilizadores em jornais *online* através da sua interação com conteúdos noticiosos. Uma vez que a coleção de dados disponível contém informações desde o ano de 2005, vamos também perceber como o tempo alterou os hábitos dos utilizadores. Vai procurar descobrir-se, por exemplo, como evoluíram os dispositivos em que os utilizadores consultam as páginas, com que expressão a página chegou a outras partes do país e do mundo, em que altura algumas redes sociais explodiram e, em sentido inverso, quando é que outras começaram o seu declínio. De forma a adaptar os conteúdos das páginas web, e até mesmo a sua estrutura, ao que os utilizadores procuram é importante reconhecer padrões de utilização para encontrar quais são os percursos que a maioria dos utilizadores segue ao visitar um jornal *online*, quais as secções do jornal mais frequentemente acedidas e quais as páginas que são visitadas com maior frequência, quando previamente se visitaram determinadas outras páginas, para procurar prever os seus próximos passos.

1.3 Estrutura da Dissertação

Para além da introdução, esta dissertação contém mais 4 capítulos. No Capítulo 2, é feita referência ao *Data Mining*, com incidência nas técnicas de *Web Mining*, especialmente *Web Usage Mining*, tendo como base alguns trabalhos relacionados com esta área e é explicada de que maneira é feita a análise de *logs*. De seguida são apresentados alguns trabalhos relacionados. Na secção final é feita a apresentação do JPN. No Capítulo 3, é apresentado o problema, juntamente com a descrição da solução usada para o resolver. São mostradas as tecnologias que contribuíram para alcançar os objetivos pretendidos. No Capítulo 4 são analisados e discutidos os resultados obtidos. No Capítulo 5 encontra-se a conclusão do trabalho e o que pode ser feito no futuro. Nos anexos é incluído um inquérito efetuado a pessoas da área do jornalismo.

Capítulo 2

Enquadramento

2.1 JornalismoPortoNet

O JPN é um jornal multimédia de atualização permanente que é, também, um projeto da Licenciatura em Ciências da Comunicação da Universidade do Porto. Acompanha a evolução das novas tecnologias de comunicação e põe em prática as mais modernas técnicas de expressão jornalística na Internet [JPNb]. Está presente na Web e em diversas redes sociais. Os seus mais de 10 anos de existência, permitem um estudo que atravessa épocas distintas, como é o caso da explosão dos dispositivos móveis ou de algumas redes sociais, bem como do desaparecimento de outras, o que tem um enorme interesse para perceber como o tempo alterou os hábitos de consumo dos utilizadores. A Figura 2.1 mostra a página de entrada do JPN, no dia 26/06/2017.

Neste capítulo vai ser feita uma breve referência às outras técnicas de Web Mining, para além de Web Usage Mining (WUM) - Web Content Mining e Web Structure Mining. De seguida, vão ser explicados os passos a seguir para a extração de informação através da análise de registos dos servidores - a metodologia de WUM, e que tipo de análise é possível fazer através desta técnica de Web Mining. No tópico seguinte serão apresentadas informações acerca da análise de registos dos servidores.

2.2 Web Mining

Web Mining é a aplicação de técnicas de *Data Mining* para extrair conhecimento de dados *Web*, que incluem documentos *Web*, hiperligações entre documentos, registos de acesso de *web-sites* [SDK05]. Etzioni [Etz96] propõe a divisão de *Web Mining* em três sub-tarefas: descoberta de recursos, extração de informação e generalização. A primeira consiste em descobrir na *Web* os documentos e serviços pretendidos. A segunda passa por extrair, dinamicamente, informação dos documentos e serviços descobertos na fase anterior. Para a última sub-tarefa são usadas técnicas

Figura 2.1: A página de entrada do JPN em 26/06/2017 [JPNa]

de *machine learning* e de *data mining* para generalizar a informação extraída. A estas sub-tarefas Kosala e Blockeel [KB00] acrescentam ainda uma análise para validar os padrões descobertos.

Tendo em conta os tipos de dados a serem analisados, podemos dividir o *Web Mining* em três categorias:

- **Web Content Mining** — Procura documentos na *web*, cujo conteúdo vá de encontro às *querys* dos utilizadores [BS02a]. Em Morianga et al. [MYTF02] é usada esta técnica de *Web Mining* para desenvolver uma *framework* que permite fazer a recolha de opiniões acerca de produtos escolhidos pelos utilizadores através da identificação de páginas web que contêm o nome desses produtos.

- **Web Structure Mining** — Analisa a estrutura de ligações da internet de forma a encontrar documentos relevantes [BS02a]. O algoritmo *PageRank*, desenvolvido por Sergey Brin e Lawrence Page [BP98], usado pela Google para classificar os resultados de uma pesquisa no seu motor de busca é exemplo do uso desta técnica de *Web Mining*. O seu funcionamento é discutido por Kumar e Singh [PS10].

- **Web Usage Mining** — Encontra padrões de utilização a partir de dados recolhidos através da interação dos utilizadores com páginas *web* para identificar o seu comportamento [Mob07].

Uma vez que o objetivo deste trabalho é analisar os utilizadores e o seu comportamento ao interagir com conteúdos noticiosos, o seu foco é o WUM, que vai ser aprofundado na Secção seguinte.

2.3 Web Usage Mining

O objetivo de WUM é encontrar padrões de utilização a partir de dados recolhidos através da interação dos utilizadores com páginas *web* para identificar o seu comportamento [Mob07]. Estes dados podem ser recolhidos a partir de três métodos diferentes [SCDT00], apresentados na Secção 2.3.1. Tal como num processo geral de *Data Mining*, vários autores [SCDT00] [JKA13] [Mob05] referem três fases principais no processo de WUM, após ser feita a recolha de dados, apresentados nas Secções 2.3.2 e 2.3.3.

Dado que a análise de *logs* é uma das formas de conhecer o comportamento dos utilizadores em páginas *web*, na Secção 2.3.4 são mostrados exemplos de registos de *logfiles* do JPN e é explicado como e que tipo de informação é possível extrair.

2.3.1 Recolha de Dados

A primeira fase de WUM é a recolha de dados. Estes dados são os registos dos pedidos feitos pelos utilizadores a servidores, que contêm as sessões desses utilizadores - considera-se uma sessão como sendo uma sequência de pedidos que um utilizador faz ao servidor. São três as principais fontes destes registos: ao nível do servidor, ao nível do cliente e ao nível de *proxy* [SCDT00]:

- **Servidor** — Grava explicitamente a navegação dos visitantes do site no *log* do servidor. Esta informação pode ser gravada no formato *Common log* ou no formato *Extended Common log* [Coo00]. Tal como Srivastava, et al. [SCDT00] referem, há vários problemas a ter em conta ao analisar os *logs* do servidor, como é o caso das visitas às páginas em cache ou a informação submetida via método POST, uma vez que nenhumas destas informações é registada no *logfile*. Esta perda de informação vai resultar em dificuldades acrescidas para identificar sessões.

- **Cliente** — A recolha de dados ao nível do cliente melhora os problemas relacionados com páginas em cache e identificação de sessões. No entanto, e uma vez que esta recolha tem como base agentes remotos ou a modificação do código fonte do *browser*, requer a cooperação dos utilizadores [SCDT00]. Shahabi et al. [SZAS97] propõem um agente remoto para a recolha de dados ao nível do cliente.

- **Proxy** — Funciona como um nível intermédio entre o cliente e o servidor, de forma a garantir segurança, controlo administrativo e serviços de *caching* [Coo00]. Tal como os servidores *web*, os servidores *proxy* possuem o seu próprio registo de acessos, que regista os pedidos de vários utilizadores para vários servidores *web* [SCDT00].

2.3.2 Pré-processamento

A fase seguinte é o pré-processamento dos dados, que lhes vai retirar o ruído que possam ter, bem como verificar dados inconsistentes ou em falta. Esta fase compreende vários passos: limpeza dos dados, identificação do utilizador e da sessão, conclusão do caminho [DK10]. Chaofeng [Cha06] efetuou um estudo onde pré-processou os *logs* da biblioteca da South-Central University for Nationalities. Nos resultados obtidos, e após o pré-processamento dos dados, das 747890 entradas, apenas 112783 se mostraram necessárias para a descoberta de padrões, o que permitiu identificar 55052 utilizadores e 57245 sessões. Num outro estudo, Kharwa et al. [KND13] iniciam o pré-processamento com 92168 entradas sendo que, após a limpeza, restaram apenas 26584. É fácil concluir que esta etapa é de grande importância pois o *logfile* é reduzido em cerca de 70% a 85% o que vai reduzir o tempo necessário para a sua análise.

De seguida é mostrado em que consiste cada uma das sub-fases do pré-processamento dos dados.

- **Limpeza dos dados** — Os registos irrelevantes são eliminados. Estes registos podem ser registos de informação de gráficos, vídeo e estilos, que são adicionados juntamente com o pedido da página, apesar do cliente ser o mesmo. São também eliminados os registos feitos por robots, que examinam uma página para extrair o seu conteúdo, bem como registos com código HTTP de falha [Cha06].

  A Figura 2.2 mostra um exemplo de três registos a eliminar de um *logfile* do JPN. Na primeira linha, a resposta ao pedido contém uma imagem. A segunda linha mostra uma resposta a um pedido feito por um robot. Na última linha a resposta veio com código de erro.

```
1  66.249.66.75 - - [17/Apr/2016:04:18:15 +0100] "GET /wp-content/uploads/2016/02/Johnny-Carson.jpg HTTP/1.1" 304 - "-" "Googlebot-Image/1.0"
2  62.138.0.25 - - [17/Apr/2016:03:30:30 +0100] "GET /robots.txt HTTP/1.1" 200 57 "-" "Mozilla/5.0 (compatible; seoscanners.net/1; +spider@seoscanners.net)"
3  2.82.1.247 - - [17/Apr/2016:03:44:47 +0100] "GET /wp-content/themes/jpnwp/js/builder.js HTTP/1.1" 404 1073 "http://jpn.up.pt/2016/04/15/queima-das-fitas-b
```

Figura 2.2: Exemplo de registos a apagar num logfile

- **Identificação do utilizador e da sessão** — Diferencia as sessões de cada utilizador a partir do *log* de acessos inicial. Losarwar e Joshi [LJ12] referem que há várias métodos para identificar utilizadores. Chaofeng [Cha06] defende que cada endereço de IP representa um utilizador; caso o endereço seja o mesmo mas o *browser* e/ou SO forem diferentes representam utilizadores diferentes; se uma página não for alcançável a partir de uma página anteriormente visitada, há dois utilizadores com o mesmo IP. Já para a identificação de sessões, Chaofeng [Cha06] defende que um novo utilizador implica uma nova sessão; se a página que encaminha para outra for nula, existe uma nova sessão; considera-se uma nova sessão após um tempo compreendido entre 25,5 e 30 minutos sem pedidos ao servidor.

- **Conclusão do caminho** — Permite descobrir o caminho de acesso completo do utilizador. Uma vez que as páginas em cache e em servidores *proxy* não são gravadas nos *logs*, é

necessário reconstruir o caminho do utilizador para identificar as páginas não guardadas [CD10].

Nesta segunda fase, uma preocupação constante é diferenciar quando um utilizador termina a sua sessão de quando deixa a página aberta no seu browser [JKA13], geralmente resolvido através do uso de *timeouts* – assume-se que a sessão terminou após um certo período de tempo sem pedidos ao servidor (entre 25,5 e 30 minutos [Cha06]). Há também que ter em atenção a *cache*: quando o utilizador retrocede a navegação, a página apresentada poderá estar em *cache*, e embora o utilizador esteja a seguir o seu caminho de navegação, o *log* vindo do servidor não apresenta estes dados, uma vez que não foi efetuado qualquer pedido. Este problema é minimizado recolhendo dados no lado do cliente, armazenados no *log* do *browser*, ou através do método conclusão do caminho [JKA13], quando os dados recolhidos no servidor estiverem a ser pré-processados. Estas formas de minimizar o problema dependem do conhecimento da estrutura do *website* e da informação referente aos *logs* do servidor.

2.3.3 Descoberta de padrões

É na terceira fase que são descobertos os padrões. Mais uma vez, esta fase é constituída por diferentes métodos: análise estatística, regras de associação, *clustering*, classificação, padrões sequenciais e modelação de dependência.

- **Análise estatística** — Permite realizar diversas análises estatísticas descritivas em diferentes variáveis (por exemplo, o número de visitas). É a técnica mais usada [SCDT00].

- **Regras de associação** — Relacionam páginas que, na maioria das vezes, são relacionadas em conjunto numa única sessão do servidor [SCDT00]. Cooley et al. [CMS97] mostram um exemplo de uma regra de associação "40% dos clientes que acederam à página com o URL /company/product1 também acederam a /company/product2". Agrawal e Srikant [AS+94] propõem dois algoritmos para encontrar regras de associação num grande *dataset*, comparando-os também com algoritmos já existentes [AIS93] [HS95].

- **Clustering** — Agrupa clientes ou itens com características semelhantes [SCDT00] o que facilita, por exemplo, o desenvolvimento de estratégias de *marketing* destinadas a um grupo específico [CMS97].

- **Classificação** — Mapeia dados em uma, ou várias, classes pré-definidas [FPSS96]. Srivastava et al. [SCDT00] exemplificam esta regra com "30% dos utilizadores que encomendaram um produto em /Product/Music têm uma idade compreendida entre 18 e 25 anos".

- **Padrões sequencais** — Encontra sequências que ocorrem de forma frequente, numa sequência com maior dimensão [MTV95]. Cooley et al. [CMS97] exemplificam com "60% dos clientes que fizeram uma encomenda em /company/product1, também encomendaram em /company/product1 nos 15 dias seguintes".

Tabela 2.1: Informação contida no *log*

Informação	**Conteúdo**
Identificação do utilizador	Quem visitou a página, normalmente o IP
Caminho da visita	Páginas do site pelas quais o utilizador passou
Intervalo de tempo	Quanto tempo o utilizador passou a navegar na página
Última página visitada	Última página em que o utilizador esteve antes de sair do site
Agente do utilizador	*Browser* de onde o utilizador fez o pedido ao servidor – tipo e versão do browser
URL	Recurso acedido pelo utilizador
Tipo de pedido	Qual o método usado, GET/POST

- **Modelação de dependência** — Procura desenhar um modelo capaz de representar dependências significativas entre as várias variáveis no domínio da *Web* [SCDT00], como encontrar características comuns entre os clientes que visitaram uma determinada página, num determinado intervalo de tempo [CMS97].

A última fase do processo é a análise dos padrões obtidos anteriormente. O objetivo desta fase é eliminar padrões e regras desnecessários do conjunto descoberto, ficando apenas com aquelas que acrescentam valor ao que já é conhecido [SCDT00].

2.3.4 Query Log Analysis

Os *logs* são ficheiros criados automaticamente e que contêm os registos de pedidos que os utilizadores fizeram ao servidor. A informação contida nos *logs* pode variar de acordo com o servidor. No entanto, as informações mais comuns são a identificação do utilizador, caminho da visita, intervalo de tempo, a última página visitada, agente do utilizador, URL do recurso acedido pelo utilizador, tipo de pedido [GMN11]. A Tabela 2.1 contém a informação sobre cada um dos campos mais comuns de uma entrada num *logfile*.

Os logs do JPN seguem o formato Combined Log Format, que combina os registos do IP do utilizador que fez o pedido, a identidade do utilizador, o nome do utilizador realizado por identificação HTTP, a data em que o pedido foi efetuado (com o dia, hora, e o fuso horário do local do pedido), o tipo de pedido e o URL, o código de resposta e o tamanho da resposta (em *bytes*) - Common Log Format - com mais dois campos, referentes à página anteriormente visitada e ao browser usado pelo cliente [Apa].

A Figura 2.3 mostra um exemplo de um registo num *logfile* do JPN.

Analisando esta entrada no *log*, vemos que o pedido foi feito pelo IP 85.138.133.248. A identidade e o nome do utilizador não estão disponíveis (verificável pelos - "nos segundos e terceiros campos). O pedido foi feito no dia 17 de Abril de 2016, às 03:44:34, no fuso horário GMT+1. O

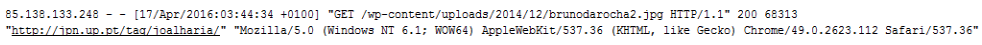

Figura 2.3: Exemplo de um registo num logfile do JPN

campo "GET /wp-content/uploads/2014/12/brunodarocha2.jpg HTTP/1.1"informa-nos que o método usado foi o GET, o endereço seguinte é o endereço da informação pedida e o HTTP/1.1 é o protocolo usado. De seguida há dois valores numéricos, sendo o primeiro o código de resposta e o segundo o tamanho do objeto retornado. O código 200 indica uma resposta com sucesso e 68313 são os bytes transferidos. Aqui termina a informação do Common Log Format [GMN11]. Os dois parâmetros que falta analisar indicam que o pedido veio de "http://jpn.up.pt/tag/joalharia/"e o browser do cliente enviou a informação "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/49.0.2623.112 Safari/537.36" [Apa].

2.4 Google Analytics

O Google Analytics[1] é uma ferramenta de *web analytics*, oferecida pela Google desde Novembro de 2005, que permite monitorizar o tráfego numa página web. Oferece estatísticas sobre diversas métricas, dimensões e conversões. Métricas são os dados que se podem medir (por exemplo, o número de páginas visualizadas ou o tempo passado no site). Dimensões são grupos de dados de utilizadores que se podem utilizar para gerar um relatório, como o tipo de dispositivo ou a localização do utilizador. As conversões são dados que indicam quantos utilizadores realizaram uma ação pretendida no site, por exemplo, a compra de um produto [Goob].

A instalação do Google Analytics para monitorizar uma página web é feita em poucos passos [Gooa]. O primeiro passo é a criação de uma conta. Após a conta ser criada é necessário adicionar código JavaScript às páginas que se quer monitorizar. Hess [Hes12] apresenta o código necessário para configurar o Google Analytics de forma a ajudar bibliotecários e programadores que mantém uma página de uma biblioteca digital a usar eventos e pesquisa no site do Google Analytics. Thushara e Ramesh [TR16] explicam o funcionamento após a implementação do código JavaScript: O utilizador faz um pedido ao servidor, que processa esse pedido e envia a resposta para o *browser*; enquanto o *browser* carrega a página, o código JavaScript do Google Analytics é executado e os dados são recolhidos (a partir do pedido, de *cookies* e de informações do *browser* e do sistema); os dados recolhidos são enviados para o servidor do Google Analytics na forma de uma imagem GIF com 1 pixel; os dados recolhidos são processados e os relatórios são criados; Os relatórios são mostrados no *dashboard* do Google Analytics.

O Google Analytics já foi usado com diversos fins. Thushara e Ramesh [TR16] usaram as suas funcionalidades para definir uma alternativa para ter sucesso no mundo do comércio eletrónico, percebendo melhor os clientes. O modelo proposto tem como objetivo seguir a sequência de cliques dos utilizadores. Os autores tiveram sucesso na identificação de tráfego e na análise

[1] https://analytics.google.com, acedido em 26/06/2017

dos utilizadores que visitam o site, tanto em diferentes países como em dois estados da Índia. Foi possível descobrir o número de sessões, número de utilizadores, número de páginas visitadas, a duração média de cada sessão e a percentagem de novos utilizadores. Omidvar et al. [OMS11] efetuaram um estudo para perceber o impacto de diferentes variáveis em variáveis dependentes (páginas visitadas dependentes do tipo de visitante - novo ou que regressou -, velocidade da conexão do visitante, e origem da visita), principalmente no número de páginas visitadas numa sessão. Foi usado um *dataset* com 19703 entradas em 23 meses. Chegaram à conclusão de que os visitantes que mais páginas visitavam era por acesso direto. No que toca ao número de páginas por velocidade de conexão, esperava-se que uma maior velocidade levasse a um maior número de páginas visitadas, mas não foi isso que se verificou. Tal deveu-se à origem geográfica da visita: o maior número de visitantes era do Irão, onde a percentagem de utilizadores com velocidade de 1.5 Mbp-s era 17%. Visitantes que retornavam à página tinham maior impacto do que os novos visitantes. Hess [Hes12] procura ajudar bibliotecários e programadores que mantêm uma página de uma biblioteca digital a usar eventos e pesquisa no site através de análise de dados fornecidos pelo Google Analytics. Foi permitido o *tracking* de eventos para recolher informação sobre cliques em links externos e download de ficheiros, bem como o *tracking* à pesquisa no site. Chegou-se à conclusão de que, em média, cada utilizador visitava 4 páginas. As páginas mais acedidas eram encontradas via links externos e a pesquisa quase não era usada. Fang [Fan07] propôs-se a usar o Google Analytics para melhorar o design e o conteúdo da página da biblioteca Rutgers-Newark Law. Para isso era preciso perceber como o site estava a ser usado, o comportamento dos utilizadores, a eficiência dos menus, fazer sugestões para melhorar a experiência dos utilizadores e, finalmente, estabelecer a melhor forma de redesenhar o *website*. Após a análise dos dados, decidiram não alterar o layout e o estilo da página. No entanto, os menus foram reorganizados: foram acrescentados, a um menu já existente, os itens mais vistos. Anteriormente estes itens, apesar de serem os mais vistos, e segundo o Google Analytics, apenas eram encontrados através de motores de busca. Pensou-se que esta solução poderia ajudar a reter utilizadores que visitassem a página pela primeira vez. Após o site ter sido redesenhado, os links para os itens mais vistos tiveram mais 30% de tráfego do que na versão anterior. Os novos visitantes aumentaram 21% e os que retornam aumentaram 44%.

2.5 Caraterização de utilizadores

Sakagami e Kamba [SK97] procuraram métodos para aprender as preferências dos utilizadores com base na sua interação com as páginas, com o objetivo de personalizar um jornal *online*. São usados uma mistura de métodos de *feedback* explícito e implícito para avaliar as preferências dos utilizadores. No *feedback* explícito os utilizadores classificam os artigos de acordo com a sua relevância. No *feedback* implícito é o sistema que infere qual foi o interesse do utilizador, com base nas ações que toma na página (por exemplo maximizar o artigo ou fazer *scroll* na sua página). Isto é possível graças a um motor de aprendizagem presente no site, que vai construindo o perfil do utilizador com base na pontuação dos artigos e nas ações tomadas pelo utilizador. Há

também um motor de *score*, que computa a importância de cada artigo comparando a frequência de ocorrência de determinadas palavras com o perfil do utilizador. O jornal foi personalizado com sucesso através dos métodos de *feedback* explícito e implícito.

Batista e Silva [BS02b] analisaram os *logs* do jornal "Público Online". Os *logs* foram adaptados aos algoritmos de *Data Mining* convertendo-os em matrizes numéricas e booleanas, onde cada coluna corresponde a uma secção do jornal, e cada linha a uma sessão. Cada matriz contém a quantidade de artigos acedidos em cada par (sessão, secção). Uma célula é verdade quando pelo menos um artigo é acedido nesse par. Uma das técnicas usadas foi a descoberta de *itemsets* frequentes (transições entre itens que ocorrem frequentemente).

Benevenuto et al. [BRCA09] analisam o comportamento das pessoas nas redes sociais usando um *dataset* construído através da análise de acessos de 37024 pessoas durante um período de 12 dias recolhido num agregador de redes sociais. Segundo os autores, estes estudos são importantes para fazer a avaliação da performance da página, permitindo a sua adaptação em termos de design ou de anúncios publicitários; para estudos sociais e marketing; para perceber o impacto das redes sociais na internet e como isso contribui para o futuro. Foram feitas 3 análises diferentes: tráfego e padrões de sessão; uma estratégia de análise para caraterizar a actividade dos utilizadores; análise da atividade dos utilizadores, através de um grafo social, no site Orkut. Foram usados 2 conjuntos de dados: *clickstream* recolhido e fornecido por um agregador de redes sociais; topologia da rede social Orkut. O estudo permitiu descobrir diferentes aspetos, alguns relacionados com a própria rede social. Os estudos prévios analisavam as interações visíveis com as redes sociais (por exemplo os comentários) - "atividades visíveis". Ao usar um *clickstream* foi possível analisar aspetos como navegar num perfil ou ver a foto de um amigo - "atividades silenciosas".

Capítulo 3

Problema e Solução

3.1 Apresentação do Problema

Com o objetivo de captar novos leitores e manter os já existentes num jornal *online*, é necessário que estes fiquem o mais satisfeitos possível ao visitar a página. Uma das maneiras de melhorar a sua satisfação é aproximar o jornal dos leitores. Uma página atrativa e com uma navegação intuitiva, ou até personalizada, é fundamental para que estes se sintam satisfeitos e se fidelizem ao jornal. *Web Analytics* é a arte de melhorar uma página web, para aumentar a sua rentabilidade, melhorando a experiência de utilização dos clientes [WK09].

Através da análise aos registos de acesso de servidores, é possível identificar padrões de utilização seguidos pelos utilizadores ao longo do tempo. É a descoberta destes padrões que vai permitir caracterizar os hábitos de utilização e navegação em jornais *online*, para perceber de que forma os utilizadores interagem com conteúdos noticiosos. O Google Analytics é uma ferramenta fornecida pela Google que, quando implementado numa página web, permite fazer, de uma maneira bastante simples, a monitorização dos utilizadores que estão, ou já estiveram, a navegar nessa página.

Neste trabalho, os padrões para estudar e caracterizar os hábitos de utilização e navegação vão ser encontrados através da análise de registos de utilização do JPN, oferecidos pelo Google Analytics.

3.2 Solução

Para descobrir os hábitos de utilização e navegação dos leitores do JPN foram seguidas várias etapas. Inicialmente foi feito um levantamento das métricas e das dimensões mais importantes, umas por escolha própria, outras em reuniões com pessoas do meio e outras através da consulta de estudos prévios na área da *Web Analytics*. Depois de escolhidas quais as métricas a usar, foi feita uma análise aos dados gerados pelo Google Analytics para perceber se era possível obter todos os

dados relevantes e, em caso afirmativo, em que intervalo de tempo estavam disponíveis - alguns dos dados não estão disponíveis desde o início - e de que maneira poderiam ser manipulados de forma à sua análise ser viável e de fácil consulta.

O passo seguinte foi a manipulação dos dados. Os ficheiros gerados pelo Google Analytics serviram de *input* a algumas ferramentas desenvolvidas em Java, que retornam novos ficheiros no formato csv. Estes novos ficheiros foram usados como *input* em *scripts*, na linguagem R, que geram os gráficos finais. A Figura 3.1 ilustra o processo seguido.

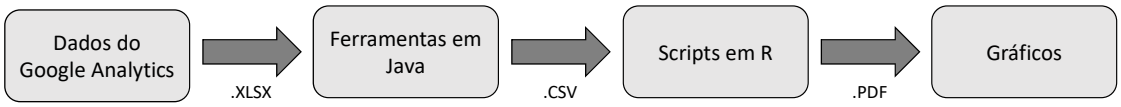

Figura 3.1: A *pipeline* de transformação dos dados

De forma paralela à recolha e processamento de dados, foi efetuado um inquérito, que circulou na redação do JPN e em estudantes dos cursos "Licenciatura em Ciências da Comunicação: Jornalismo, Assessoria, Multimédia"e "Mestrado em Ciências da Comunicação", da Faculdade de Letras da Universidade do Porto, com o objetivo de perceber, numa amostra de maior dimensão, de que maneira algumas das métricas e dimensões eram importantes para quem está na área do jornalismo.

3.2.1 Inquéritos

Os inquéritos foram efetuados com o objetivo de perceber a importância de cada uma das métricas e dimensões para quem está na área do jornalismo e se, à medida que os anos de trabalho aumentam, se verifica alguma diferença na importância que é dada às métricas e dimensões. Assim, o inquérito contém uma parte inicial onde os inquiridos indicam se são estudantes ou profissionais e o ano de curso, no caso de estudantes, ou o número de anos de experiência, no caso de profissionais.

Após a secção de caraterização do inquirido, o questionário encontra-se dividido em três secções, que correspondem a três níveis na organização da página web: artigos, categorias e página web. A secção correspondente aos artigos permite classificar uma série de métricas e dimensões de acordo com o que se considera relevante para compreender o impacto do artigo junto dos leitores. A secção correspondente às categorias permite classificar as métricas e dimensões de acordo com o que se considera relevante para compreender o impacto da página que apresenta os artigos de uma categoria. Finalmente, a secção correspondente à página web permite classificar as métricas e dimensões, de acordo com o que se considera relevante, tendo em conta todas as visitas ao jornal. Em todas as secções há a possibilidade do inquirido sugerir novas métricas, através de um campo que permite uma resposta aberta.

Tabela 3.1: Dados recolhidos no Google Analytics

Métrica / Dimensão	**Período**	**Fonte**
Número de utilizadores	Novembro de 2005 a Abril de 2017	[TR16]
Número de sessões	Novembro de 2005 a Abril de 2017	[TR16]
Número de sessões/utilizador	Novembro de 2005 a Abril de 2017	[TR16]
Número de páginas visitadas	Novembro de 2005 a Abril de 2017	[OMS11]
Número de páginas visitadas/sessão	Novembro de 2005 a Abril de 2017	[OMS11]
Idade dos utilizadores	Janeiro de 2015 a Abril de 2017	Reuniões
Dispositivos usados	Dezembro de 2005 a Abril de 2017	[TR16]
Número de visitas/hora	Novembro de 2005 a Abril de 2017	Reuniões
Sexo dos utilizadores	Janeiro de 2015 a Abril de 2017	Reuniões
Número de visitas/dia da semana	Novembro de 2005 a Abril de 2017	Reuniões
Tempo de leitura	Novembro de 2005 a Abril de 2017	[TR16]
Localização geográfica (Portugal e Mundo)	Novembro de 2005 a Abril de 2017	[TR16]
Palavras mais pesquisadas	Abril de 2016 a Abril de 2017	[Fan07]
Visitas com pesquisa	Abril de 2016 a Abril de 2017	[Fan07]
Fidelização (percentagem de novas sessões)	Outubro de 2008 a Abril de 2017	Reuniões
Redes Sociais	Dezembro de 2005 a Abril de 2017	Reuniões
Origem da visita	Novembro de 2005 a Abril de 2017	[TR16]
Percentagem de novas sessões	Dezembro de 2005 a Abril de 2017	[TR16]

3.2.2 Dados do Google Analytics

O Google Analytics permite a recolha de diversos tipos de dados. No entanto, nem todos os dados fornecidos são relevantes para este estudo, nem a data a partir da qual estes dados estão disponíveis é a mesma. Para sistematizar os dados a estudar e o intervalo de tempo em que estes dados estão disponíveis foi construída a tabela 3.1 que, para além destas informações, apresenta a fonte que permitiu o estudo da métrica/dimensão em questão.

3.2.3 Compilação dos Dados

A compilação dos dados fornecidos pelo Google Analytics foi feita em Java, através da utilização da biblioteca Apache POI, cujo funcionamento é explicado na secção 3.3.2. Os dados do Google Analytics são extraídos no formato Excel. As ferramentas desenvolvidas com recurso à biblioteca Apache POI têm duas finalidades distintas: juntar os dados de vários ficheiros Excel (esta funcionalidade não é necessária para todos os ficheiros) e selecionar os dados importantes para análise, gerando um ficheiro CSV.

Os ficheiros Excel gerados pelo Google Analytics contêm diversos dados sobre a métrica que se pretende estudar. Nem todos são necessários para a elaboração dos gráficos a estudar, pelo que é necessário filtrar os dados necessários. A Figura 3.2 mostra partes de cada uma das 3 folhas incluídas no ficheiro Excel correspondente às idades dos visitantes entre o dia 1 de Janeiro de 2015 e 30 de Abril de 2017. Estas datas são consultadas na folha de resumo, Figura 3.2a. As duas páginas seguintes são as páginas que mostram os valores referentes às métricas em estudo.

JPN
Dados demográficos: Idade
20150101-20170430
Links para dados:
Conjunto de Dados1
Conjunto de Dados2

(a) Folha 1 - Resumo

Idade	Sessões	% de novas sessões	Novos Utilizadores	Taxa de rejeições	Páginas/Sessão	Duração média da sessão	Taxa de conversão de objetivos	Objetivos Concluídos	Valor do Objetivo
25-34	254691	78,83%	200763	85,61%	1,35	55,55	0,00%	0	0,00
18-24	203078	74,79%	151883	84,51%	1,41	66,42	0,00%	0	0,00
35-44	173878	83,22%	144707	86,60%	1,28	41,92	0,00%	0	0,00
45-54	94065	82,37%	77482	85,25%	1,31	44,22	0,00%	0	0,00
55-64	64656	82,68%	53455	84,85%	1,32	44,59	0,00%	0	0,00
65+	29873	82,13%	24535	84,25%	1,34	50,22	0,00%	0	0,00
	820241	79,59%	652825	85,39%	1,34	52,99	0,00%	0	0,00

(b) Folha 2 - Conjunto de Dados 1

	Total	0000	0001	0002	0003	0004	0005	0006	0007	0008	0009	0010	0011	0012	0013	0014	0015
Idade	Sessões	Sessões	Sessões	Sessões	Sessões	Sessões	Sessões	Sessões	Sessões	Sessões	Sessões	Sessões	Sessões	Sessões	Sessões	Sessões	Sessões
25-34	254691	3751	5557	10994	8281	9356	7450	6022	4729	5214	6453	6717	4807	6484	6229	9828	15014
18-24	203078	2159	3516	7158	6142	6751	5075	4155	3470	4557	5693	7030	4640	6193	5967	9502	15136
35-44	173878	1830	3210	6043	4966	6212	5586	4733	3606	4121	4775	4890	3895	5023	4504	7645	10277
45-54	94065	958	1738	3358	2774	3400	2704	2545	1868	2196	2626	2688	2040	2861	2610	4207	5604
55-64	64656	602	1134	2212	1748	2257	1859	1663	1377	1523	1596	1728	1406	1984	1802	2644	3999
65+	29873	303	671	1147	1003	1516	966	776	543	625	667	725	587	851	712	1127	1571
	820241	9603	15826	30912	24914	29492	23640	19894	15593	18236	21810	23778	17375	23396	21824	34953	51601

(c) Folha 3 - Conjunto de Dados 2

Figura 3.2: As 3 folhas que constituem o ficheiro Excel gerado pelo Google Analytics para a idade dos visitantes

O primeiro conjunto de dados, Figura 3.2b, contém várias informações para cada grupo. Os dados necessários para o desenho dos gráficos estão na última folha, mostrada na Figura 3.2c. São os números desta última folha que vão ser reformatados e guardados num ficheiro CSV que vai servir, posteriormente, como *input* a um *script* escrito em R. As primeiras linhas do ficheiro CSV gerado são apresentadas na Figura 3.3.

A outra ferramenta desenvolvida agrupa o número de visitas de acordo com o desejado. Esta ferramenta foi usada para se obter a variação da percentagem de sessões ao longo do dia e ao longo da semana. O Google Analytics permite obter o número de sessões em função da hora, do dia, da semana ou do mês. Para perceber como as visitas variam ao longo do dia, foi extraído o número de sessões em função da hora, o que resulta num ficheiro XLSX com cerca de 100000 linhas, de Novembro de 2005 a Abril de 2017. O total de sessões em função da hora é obtido somando os valores de cada linha correspondente a essa hora, sendo posteriormente convertidos em percentagens.

A automatização destes processos com a criação e utilização destas duas ferramentas permitiu uma grande economia de tempo, uma vez que este processo se tornaria bastante moroso caso fosse feito à mão.

id	25-34	18-24	35-44	45-54	55-64	65+
1	3751	2159	1830	958	602	303
2	5557	3516	3210	1738	1134	671
3	10994	7158	6043	3358	2212	1147
4	8281	6142	4966	2774	1748	1003
5	9356	6751	6212	3400	2257	1516
6	7450	5075	5586	2704	1859	966
7	6022	4155	4733	2545	1663	776
8	4729	3470	3606	1868	1377	543
9	5214	4557	4121	2196	1523	625
10	6453	5693	4775	2626	1596	667

Figura 3.3: Primeiras linhas do ficheiro CSV gerado pela ferramenta desenvolvida em Java

3.3 Tecnologias

3.3.1 R

R[1] é uma linguagem e um ambiente para computação estatística e produção de gráficos, que permite a manipulação e armazenamento de dados de forma eficaz. Inclui um conjunto de operadores que permitem efetuar cálculos em vetores, uma coleção de ferramentas para análise de dados, bem como ferramentas que permitem analisar e exibir gráficos [Pro]. Esta linguagem é facilmente extensível através do uso de *packages*. Esta linguagem foi usada para o desenvolvimento de gráficos, através do desenvolvimento de *scripts* que fazem a leitura de um ficheiro CSV e geram o gráfico pretendido. O *package* usado nos *scripts* para o desenvolvimento dos gráficos foi o ggplot2.

Ggplot2[2] é um sistema de desenho para a linguagem R, baseado na gramática dos gráficos - uma ferramenta que permite descrever de forma concisa as componentes de um gráfico, como é o caso do *dataset*, da escala e do sistema de coordenadas [Wic10] - que usa diversos componentes existentes em outros *packages* para desenho de gráficos [Wic13], como é o caso do *base* [Tcw] e do *lattice* [Sar].

3.3.2 Apache POI

Apache POI[3] é uma biblioteca para Java que permite manipular ficheiros com base nos padrões OOXML (XLSX, DOCX, PPTX) e no formato Microsoft's OLE 2 Compound Document (XLS, DOC, PPT) [Foub]. Com o uso desta biblioteca é possível ler e escrever ficheiros excel a partir de um programa em Java. A missão da Apache POI Project é ter uma API para a manipulação de ficheiros com base nos padrões OOXML e OLE2 [Foub]. Para a elaboração das ferramentas foi usada a implementação da Apache POI para o formato XLSX em Java puro - XSSF. Esta implementação fornece diversas formas de ler, criar e modificar folhas de cálculo XLSX [Fouc]. Diversas pessoas, empresas e serviços usam a Apache POI para cumprir os seus objetivos. É o caso, por exemplo, do Bank of Lithuania, que usa a API HSSF para os seus dados estatísticos,

[1] https://www.r-project.org/, acedido em 26/06/2017
[2] http://ggplot2.org/, acedido em 26/06/2017
[3] https://poi.apache.org/, acedido em 26/06/2017

ou da Deutsche Bahn, que usa a componente HWPF para processar documentos de especificações complexos armazenados no antigo formato do Microsoft Word [Foua].

Capítulo 4

Discussão dos Resultados

4.1 Visitas e sessões

O JPN começou a ser monitorizado pelo Google Analytics em Novembro de 2005. No entanto, e como já foi referido, algumas das métricas não estão disponíveis desde essa altura pelo que, dependendo da métrica a estudar, a data de início varia tendo em conta a data em que começou a ser monitorizada. Os dados analisados compreendem um período temporal entre essa data e o dia 30 de Abril de 2017. Entre o início de Novembro de 2005 e o fim de Abril de 2017, o JPN teve um total de 19719 artigos, dos quais 1304 foram artigos multimédia. A Figura 4.1 mostra o número de artigos e artigos multimédia publicados em cada mês e a Figura 4.2 mostra o número total de artigos publicados em cada mês.

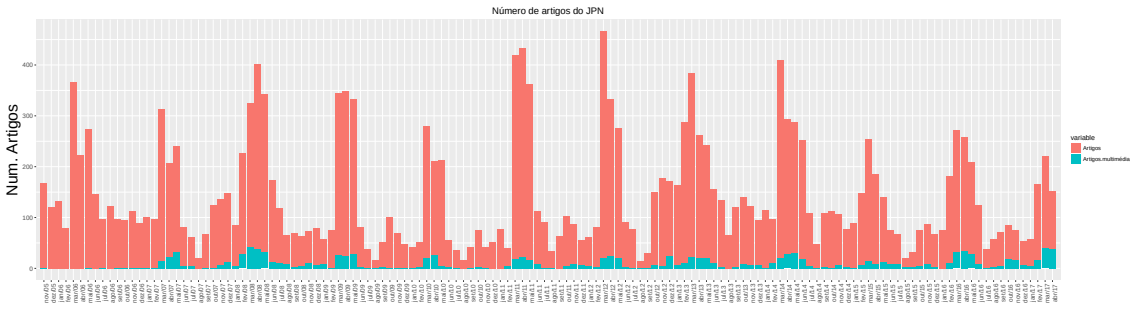

Figura 4.1: Número de artigos e artigos multimédia

Estes artigos encontram-se divididos em 7 principais categorias, presentes numa barra na página de entrada: UP, Academia, Grande Porto, Portugal, Cultura, Desporto, Mundo. As categorias Portugal, Grande Porto e Cultura são as que mais artigos têm. A Figura 4.3 mostra o número de artigos publicados em cada categoria.

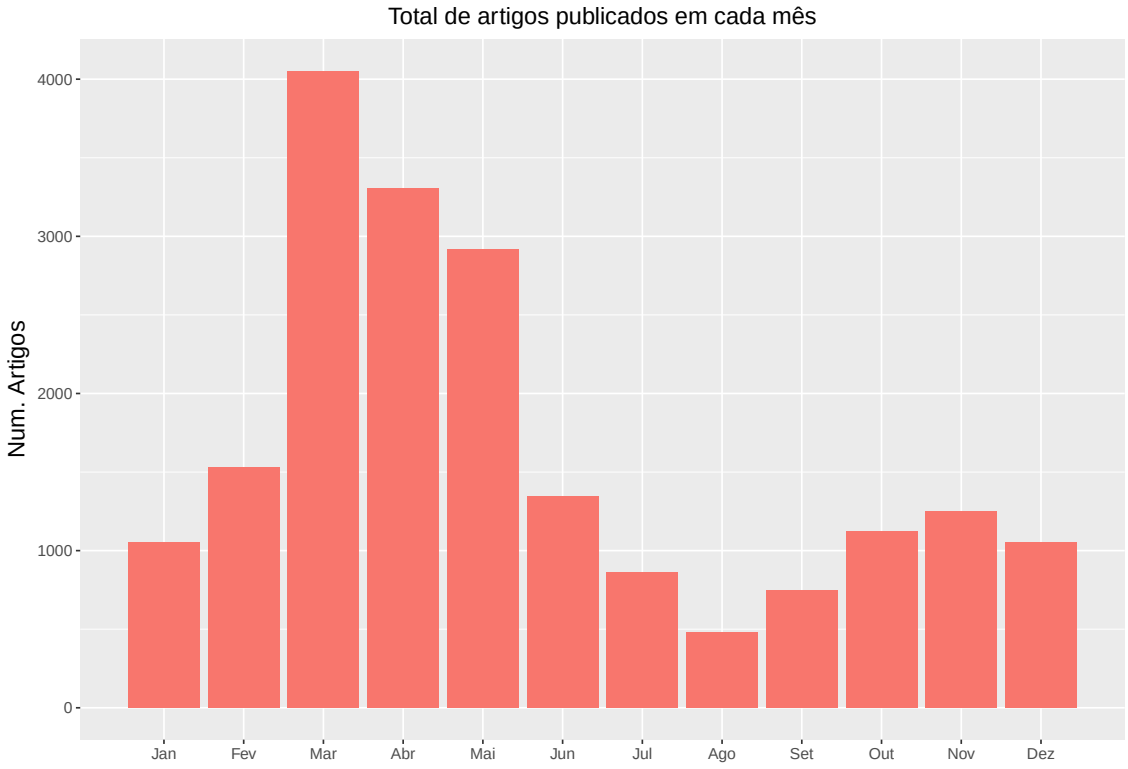

Figura 4.2: Número total de artigos publicados em cada mês

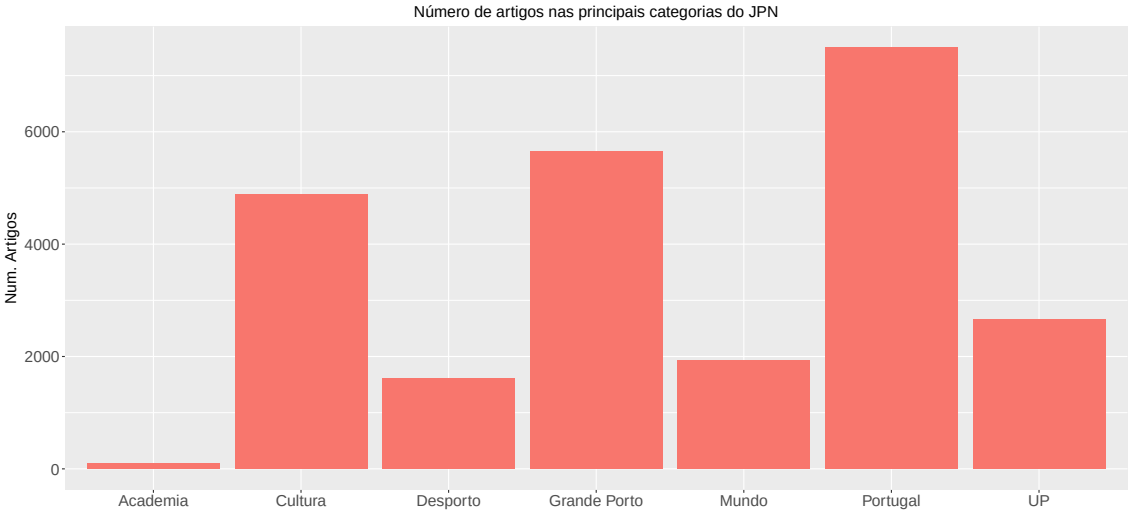

Figura 4.3: Número de artigos nas principais categorias

Quanto aos visitantes foram 6.691.792, que efetuaram 8.344.751 sessões. A primeira análise efetuada serviu para perceber se, ao longo do ano, havia alguma alteração nos acessos dos utilizadores às páginas do JPN. É facilmente identificável um padrão: os picos do gráfico correspondem

aos meses de Março e Maio, havendo depois uma descida no número de visitantes até às depressões que correspondem aos meses de Verão, principalmente o mês de Agosto. A partir desta altura os números voltam a aumentar, normalmente com uma quebra no mês de Dezembro. Por aqui se nota o ambiente académico em que este jornal se insere, uma vez que mais utilizadores correspondem ao final do ano letivo e menos utilizadores correspondem aos períodos sem aulas. As Figuras 4.4 e 4.5 mostram a distribuição dos utilizadores e das sessões ao longo dos meses.

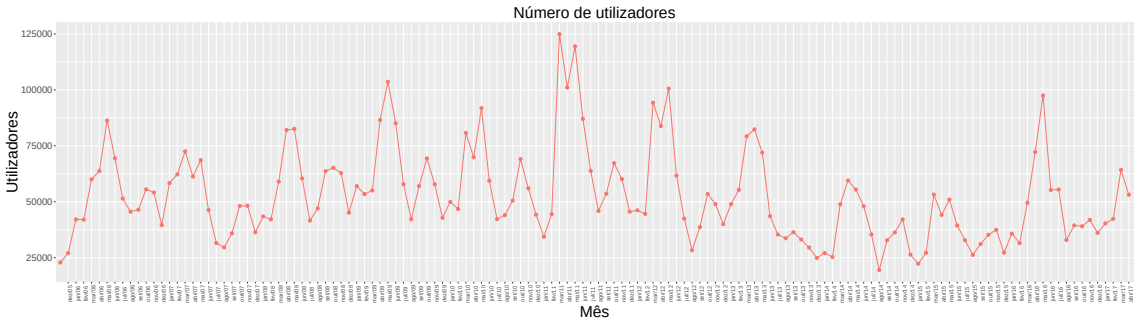

Figura 4.4: Distribuição dos utilizadores ao longo dos meses

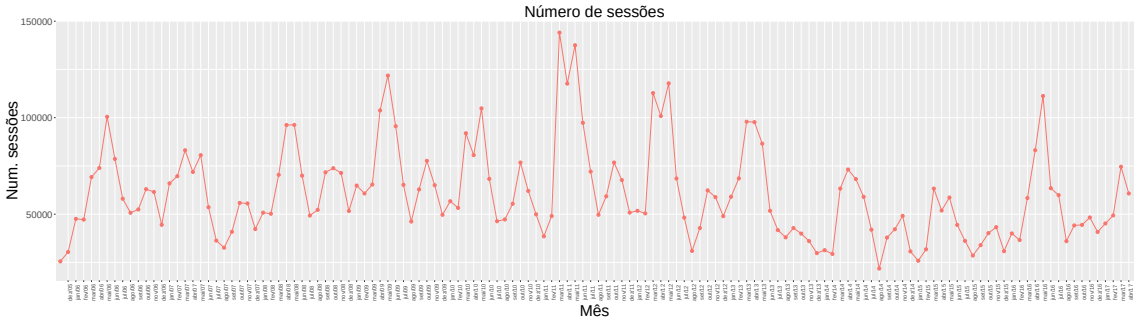

Figura 4.5: Distribuição das sessões ao longo dos meses

Olhando para o gráfico dos artigos e das sessões, vê-se que as oscilações são, na maioria dos meses, semelhantes: quanto mais artigos forem publicados num mês, mais sessões são efetuadas. A Figura 4.6 mostra a semelhança na variação do número de sessões e de artigos publicados, em função do mês.

Com o número de utilizadores e o número de sessões tão aproximados, a quantidade de sessões por utilizador é, naturalmente, reduzida. Entre Novembro de 2005 e Abril de 2017, a média de sessões por utilizador é de 1,15, um número que mostra que a percentagem de novas sessões - isto é, a primeira vez que um utilizador acede às páginas do JPN - é sempre bastante elevado, quase sempre acima dos 75%. A Figura 4.7 mostra a distribuição do número de sessões por utilizador e a Figura 4.8 mostra a distribuição da percentagem de novas sessões ao longo dos meses.

Ainda no mesmo período, o número total de páginas visitadas foi de 13.237.625, uma média de 1,59 páginas em cada sessão. A Figura 4.9 mostra a distribuição do número de páginas visitadas ao longo dos meses e a Figura 4.10 mostra a distribuição da média de páginas visitadas por sessão

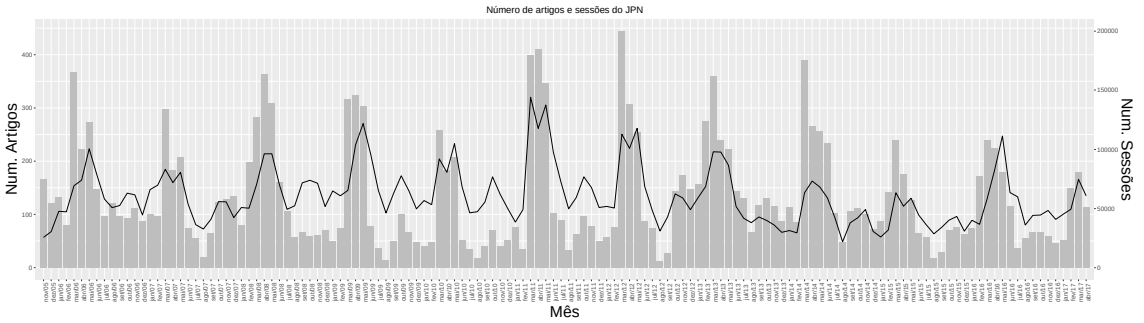

Figura 4.6: Distribuição do número de artigos e sessões ao longo dos meses

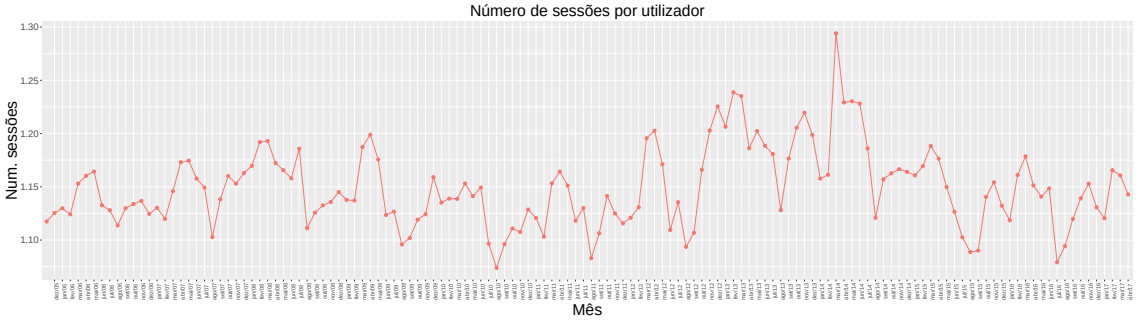

Figura 4.7: Distribuição das sessões por utilizador ao longo dos meses

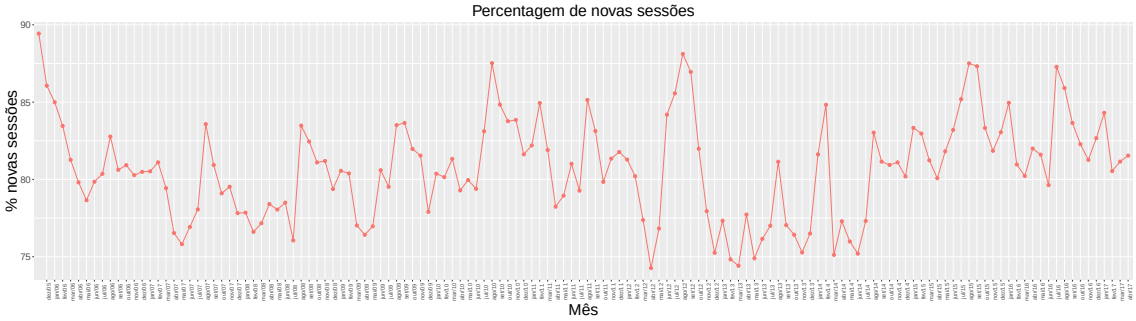

Figura 4.8: Distribuição da percentagem de novas sessões ao longo dos meses

ao longo dos meses. O número total de páginas visitadas em cada mês varia da mesma maneira que o número de utilizadores e o número de sessões.

4.2 Percurso dos utilizadores

A partir de Outubro de 2008 é possível analisar o percurso dos utilizadores nas páginas do JPN. No período que compreende os meses de Outubro de 2008 a Abril de 2017, a página mais acedida em primeiro lugar foi a página que mostra todas as notícias do ano de 2007. Uma vez que o número de páginas visualizadas em cada sessão é muito baixo, as desistências (não navegar

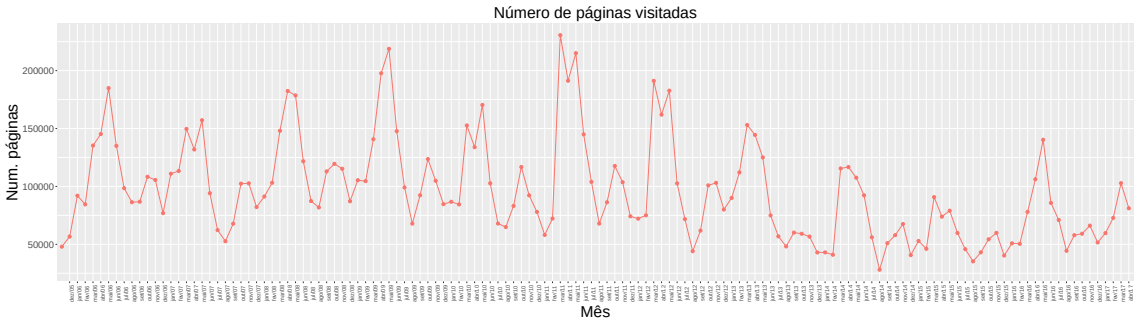

Figura 4.9: Distribuição do número de páginas visitadas ao longo dos meses

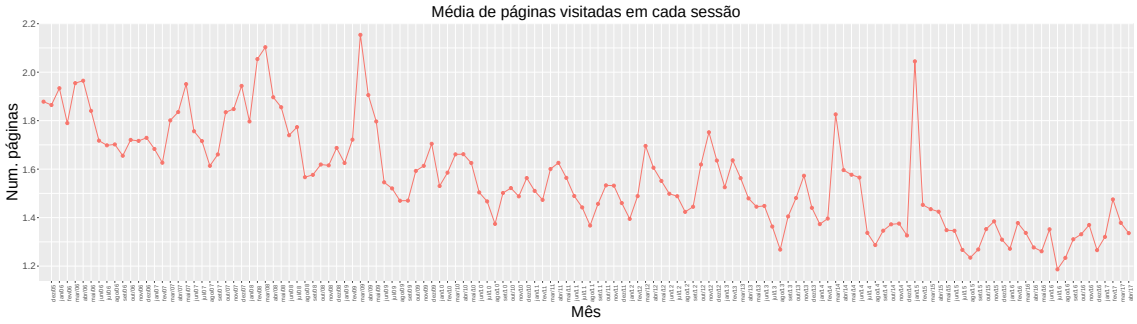

Figura 4.10: Distribuição da média de páginas visitadas por sessão ao longo dos meses

para outra página) são muito elevadas. No caso da página mais visitada, apenas 6,05% das sessões passou para uma nova página. Este número é mais baixo na página de entrada. 315.412 sessões tiveram a página de entrada como primeira página visitada, das quais 148.699 não passaram para outra página. Este número permite concluir que 53% dos utilizadores que iniciaram sessão na página de entrada visitaram, pelo menos, duas páginas, um número superior à média de páginas por sessão, uma vez que apenas 8,8% das sessões teve mais de uma página visitada. A Figura 4.11 compara a percentagem de sessões que iniciaram na página de entrada e tiveram mais de uma página visitada com a percentagem de sessões total com mais de uma página visitada.

4.3 Demografia

Os dados demográficos dos visitantes do JPN apenas estão disponíveis a partir de Janeiro de 2015. No período compreendido entre esta data e o final de Abril de 2017, o grupo etário entre os 25 e os 34 anos foi quem mais acedeu às páginas do jornal, com 253.495 sessões, o que corresponde a 31,15% do número total de sessões - 813.907. Logo a seguir, com 206.389 sessões (correspondente a 25,36% do número total de sessões) está a faixa etária entre os 18 e os 24 anos. As Figuras 4.12 e 4.13 mostram a distribuição das visitas por idade em função do mês e a distribuição da percentagem de visitas por idade em função do mês. É aqui notório, uma vez mais, o ambiente académico em que este jornal está inserido. Em termos de sexo, as mulheres

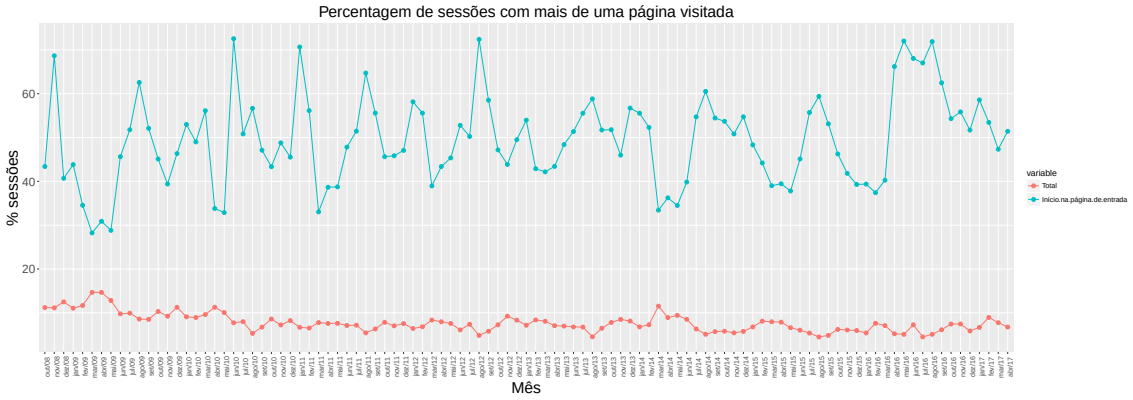

Figura 4.11: Percentagem de sessões com mais de uma página visitada

efetuam mais sessões (57,01%) do que os homens (42,99%), com uma diferença de mais 4587 sessões por mês.

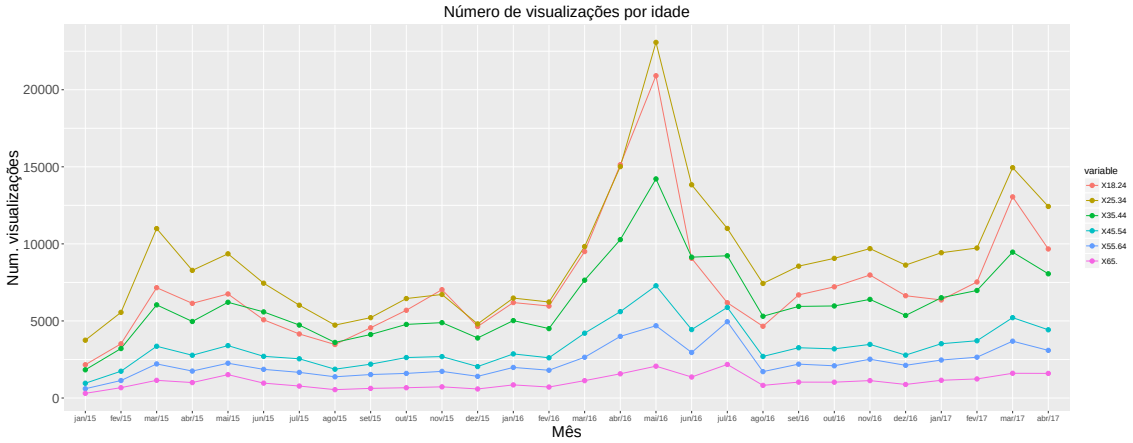

Figura 4.12: Distribuição das visitas por idade em função do mês

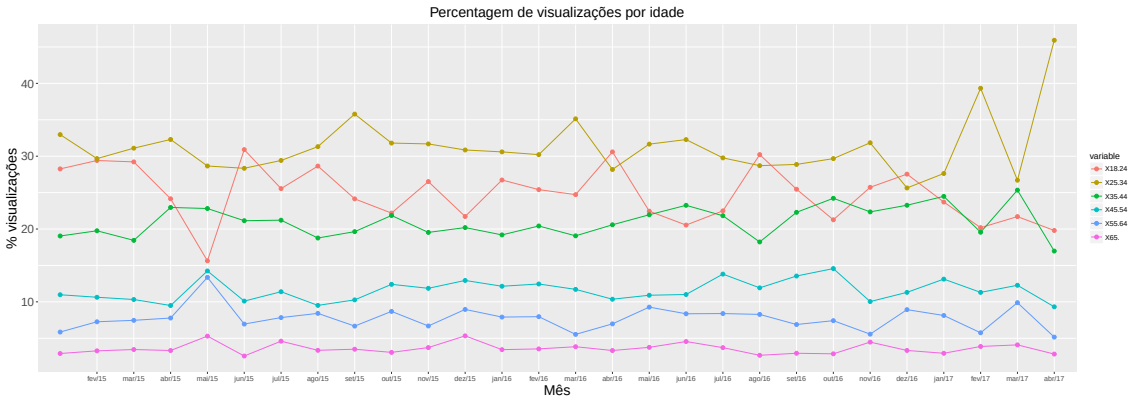

Figura 4.13: Distribuição da percentagem de visitas por idade em função do mês

4.4 Geografia

A maioria das sessões do JPN têm origem em Portugal (6.264.126 - 75,07% do total de sessões), nomeadamente no distrito do Porto (29,3% do total de sessões) e de Lisboa (26,2% do total de sessões). Os distritos de Aveiro, Braga e Setúbal ocupam os restantes lugares do top 5 de origem das visitas em Portugal, com 736.449 sessões - 8,83% do total de sessões. A Figura 4.14 apresenta o número de sessões nos 10 distritos que originam mais visitas ao JPN e a Figura 4.15 oculta os números dos distritos do Porto e de Lisboa, para uma melhor perceção dos restantes distritos.

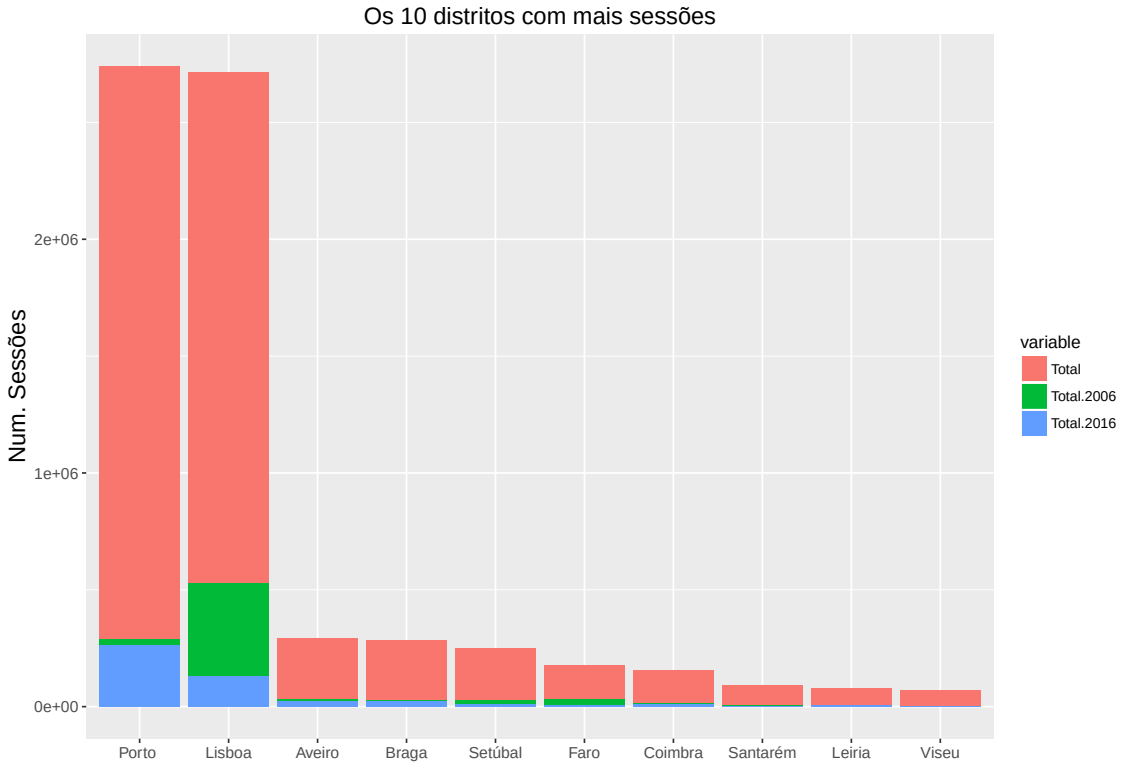

Figura 4.14: Os 10 distritos com mais sessões

No resto do Mundo, o Brasil é o segundo país com maior número de sessões (1.470.980 - 17,63% do total de sessões). França, Espanha e Estados Unidos contribuíram com 187.347 sessões - 2,25% do total de sessões. A Figura 4.16 apresenta o número de sessões nos 10 países que originam mais visitas ao JPN e a Figura 4.17 oculta os números de Portugal e Brasil, para uma melhor perceção dos restantes países. A origem das sessões a partir de Portugal não tem influência na primeira página visitada, uma vez que a maioria dos acessos têm origem na página de entrada. Tal como no estudo de Fang [Fan07], de uma página direcionada para o meio académico, a maioria dos acessos é feita no país e região de origem da página.

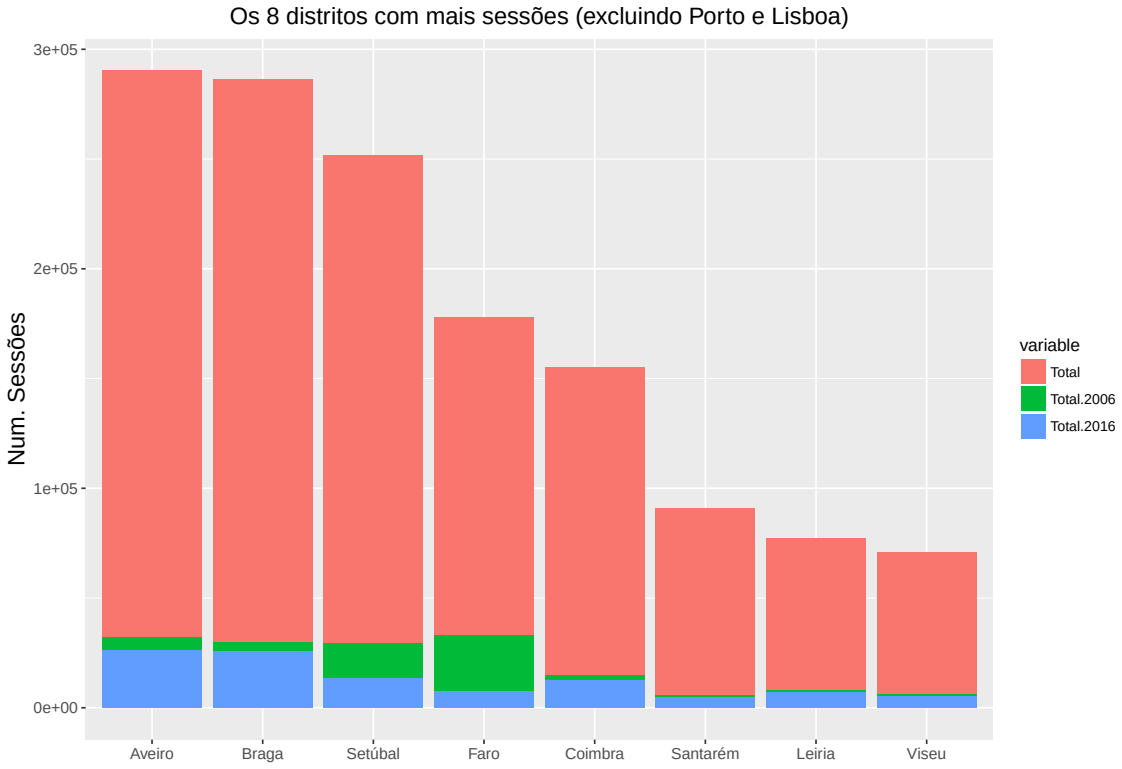

Figura 4.15: Os 8 distritos com mais sessões (excluindo Porto e Lisboa)

4.5 Dispositivos usados nos acessos ao JPN

O grande intervalo em que muitos dos dados estão disponíveis é particularmente interessante para confirmar certas tendências notórias no dia-a-dia. É o caso da utilização de dispositivos móveis e da utilização de algumas redes sociais em detrimento de outras. Os dados referentes ao tipo de dispositivo usado foram recolhidos anualmente. Os primeiros registos de acessos a partir de telemóveis são do ano de 2009, enquanto que os acessos com *tablets* são do ano de 2011. É notória a diminuição dos acessos a partir de computador e o aumento dos acessos a partir de dispositivos móveis, principalmente telemóveis. A Figura 4.18 mostra o número de sessões em função do ano e do tipo de dispositivo usado, enquanto que a figura 4.19 mostra as percentagens dos acessos entre *desktops*, telemóveis e *tablets*. A razão entre os acessos por dispositivos móveis em relação aos acessos por computador era de 0,017 em 2011, enquanto que nos primeiros quatro meses de 2017 foi de 0,969 o significa que, hoje em dia, já há quase tantos acessos a partir de dispositivos móveis como a partir de computadores. Este rácio significa que, caso as páginas não estivessem direcionadas para os dispositivos móveis, a necessidade de uma mudança teria de ser pensada de forma urgente. A Figura 4.20 mostra como as percentagens de acessos entre *desktops* e dispositivos móveis se têm vindo a aproximar. É expectável que, a muito curto prazo, haja mais acessos por dispositivos móveis do que por computador.

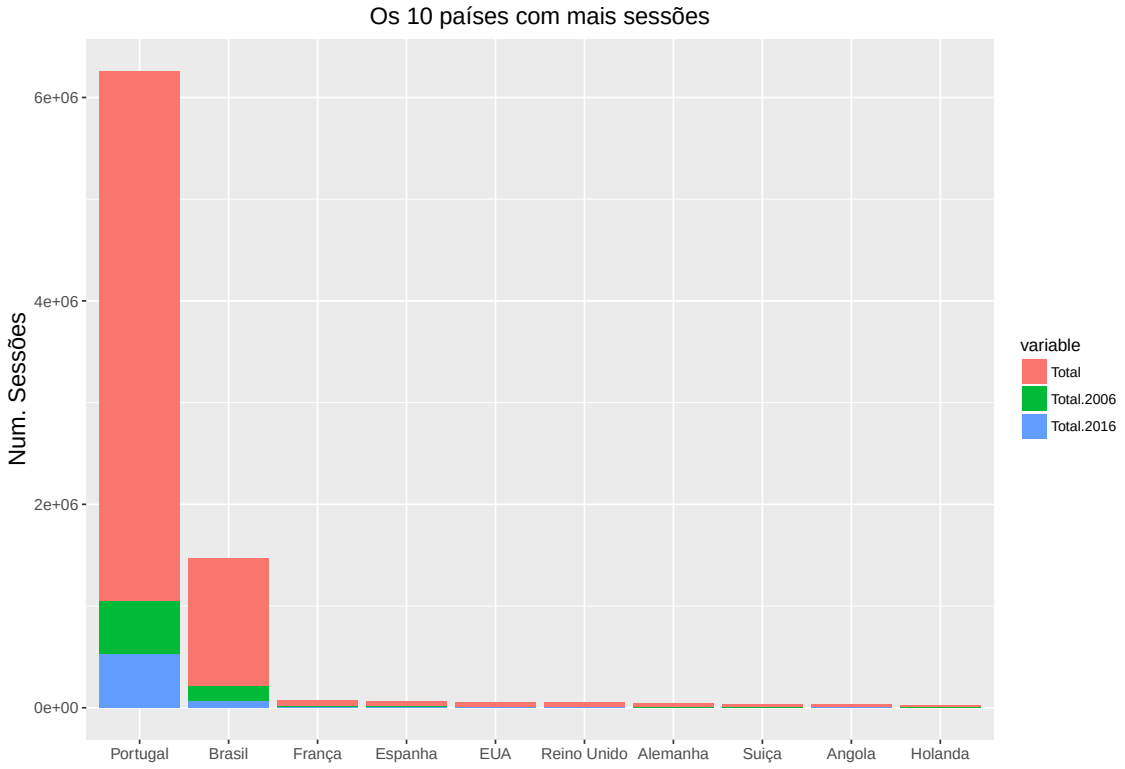

Figura 4.16: Os 10 países com mais sessões

4.6 Origem das visitas

Tal como os dispositivos usados para aceder ao jornal, as origens das visitas também têm grande importância para os editores e jornalistas, uma vez que mostram onde é preciso haver maior promoção ao jornal. O maior número de sessões tem tido origem a partir de pesquisas em motores de busca, exceção para os anos de 2013 e 2014, que foi através de referências em outras páginas. Os acessos diretos vão oscilando, mas sempre com números reduzidos. O seu máximo foi atingido em 2014, com 83.602 sessões - 10% dos acessos totais. Omidvar et al. [OMS11] referem que a maioria dos visitantes diretos são pessoas que retornam à página. Se tivermos este facto em conta, bem como o de que a percentagem de novas sessões tem sido bastante elevada, encontra-se uma justificação para a baixa percentagem de acessos diretos, o que vai de encontro aos resultados desse estudo. Já os acessos a partir de redes sociais têm tido, como seria expectável, um aumento considerável. Até 2010 foi sempre o meio de acesso menos utilizado, tendo em 2011 ultrapassado os acessos diretos e em 2015 os acessos por referência em outras páginas. Este aumento é de quase 23% em relação ao total de acessos - em 2005 0,34% das sessões tiveram origem em redes sociais, enquanto que nos primeiros quatro meses de 2017, este valor é de 22,8%. A Figura 4.21 mostra o número de sessões efetuadas em função do ano, tendo em conta a forma como se chegou às páginas do JPN e a Figura 4.22 mostra a percentagem de sessões efetuadas em função do ano, tendo em conta a forma como se chegou às páginas do JPN. Omidvar et al. [OMS11] referem que visitantes

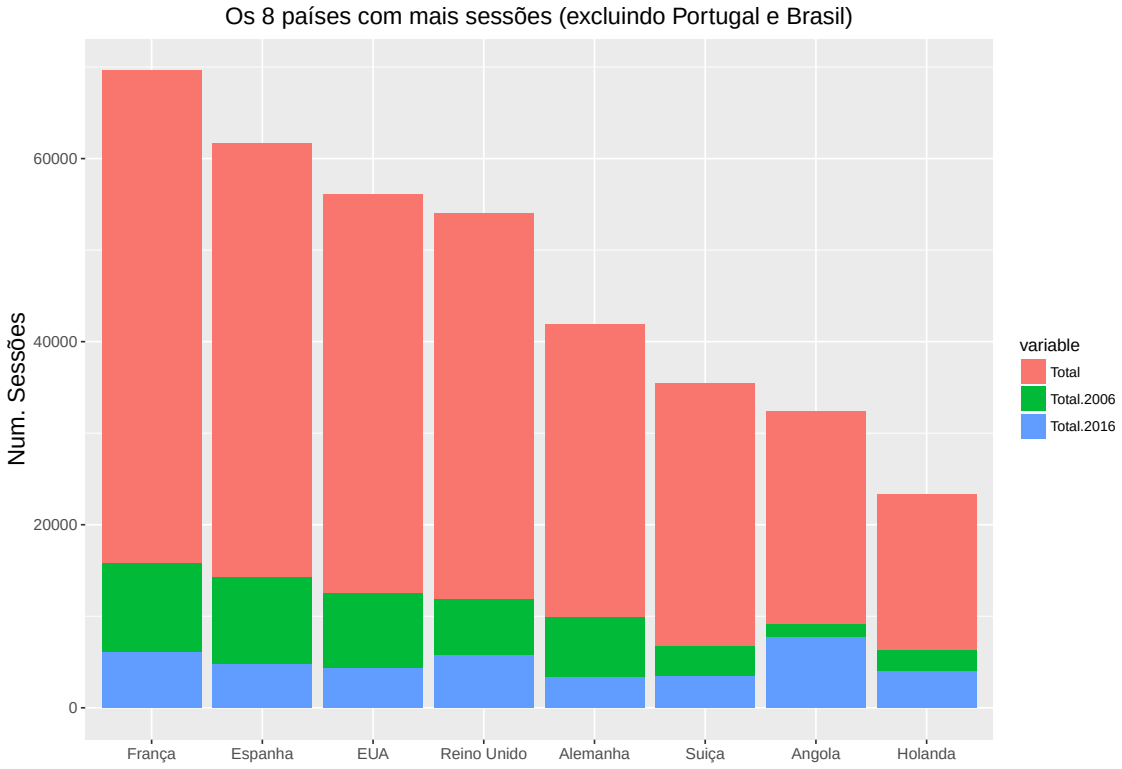

Figura 4.17: Os 8 países com mais sessões (excluindo Portugal e Brasil)

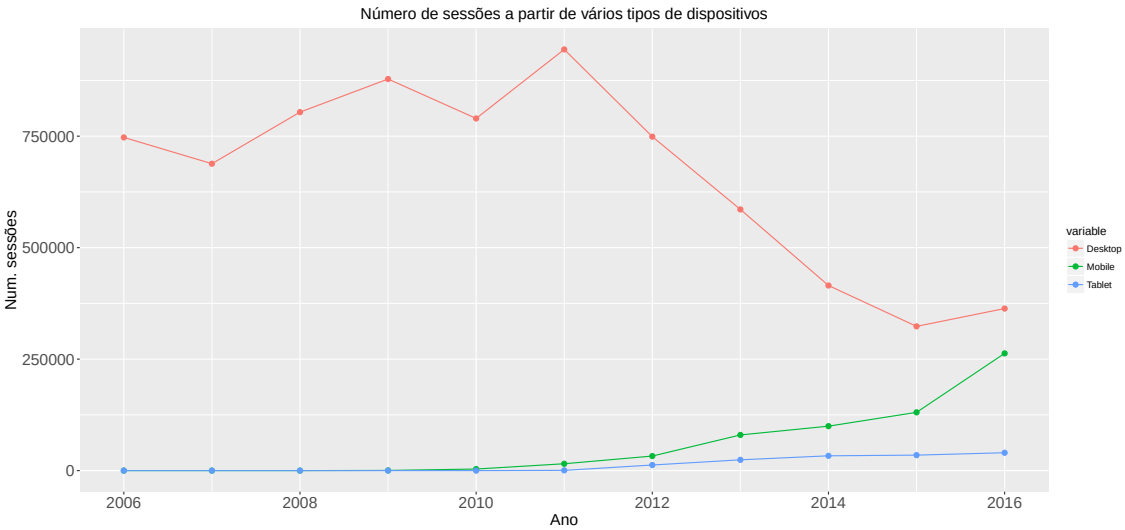

Figura 4.18: Distribuição do número de sessões a partir de vários tipos de dispositivos em função do ano

que acedem diretamente a uma das páginas acabam por visitar mais páginas do que aqueles que acedem de outra maneira. No entanto, ao analisar estes dados, a conclusão obtida não é a mesma, uma vez que o número médio de páginas visitadas é muito semelhante, independentemente da

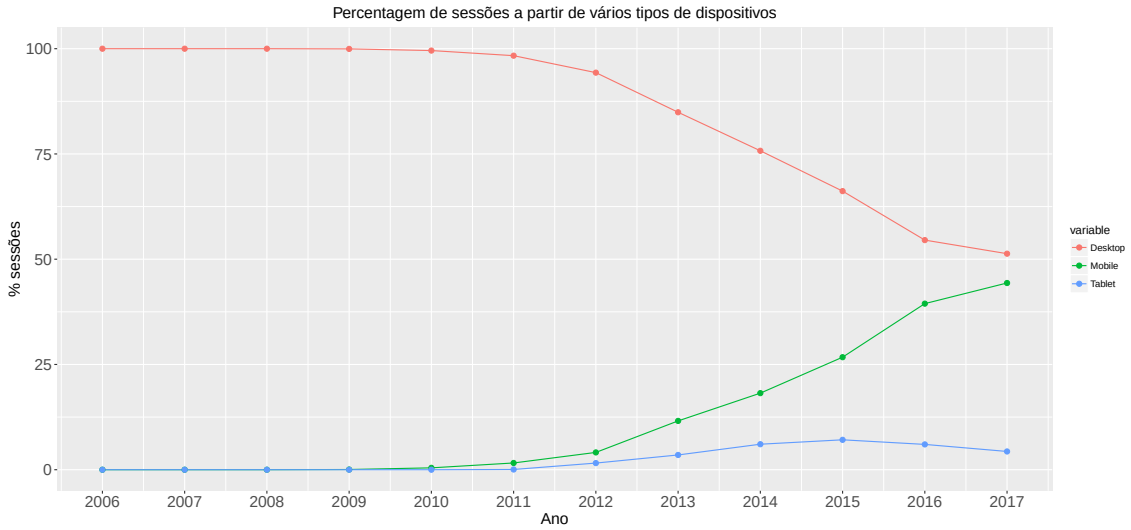

Figura 4.19: Distribuição da percentagem de sessões a partir de vários tipos de dispositivos em função do ano

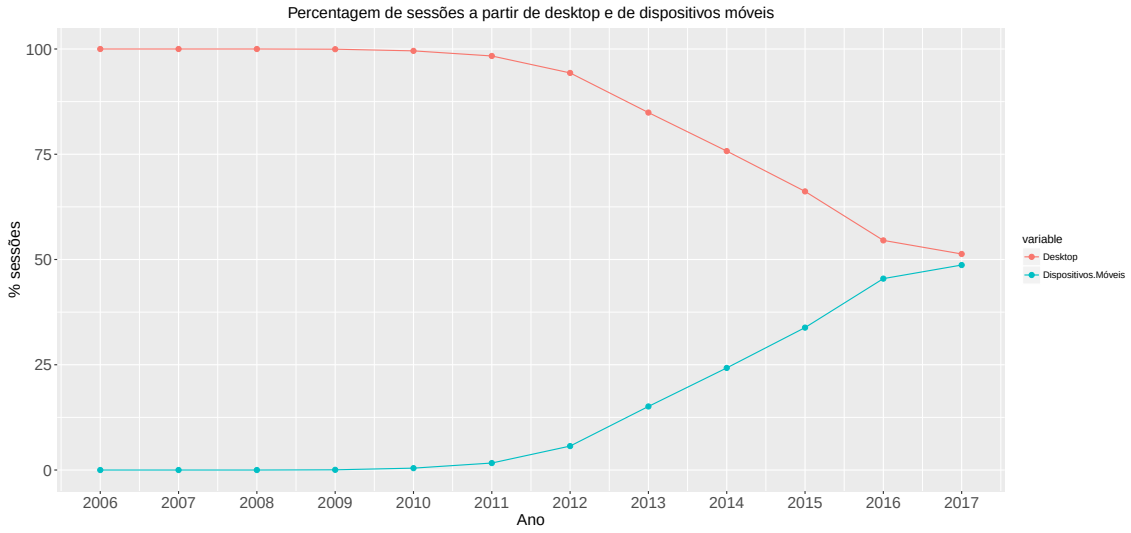

Figura 4.20: Distribuição da percentagem de sessões a partir de desktop e de dispositivos móveis em função do ano

origem da visita: 1,33 páginas para acessos via pesquisa orgânica, 1,37 páginas para acessos via redes sociais, 1,43 páginas por referência em outras páginas e 1,31 páginas por acesso direto.

4.7 Redes Sociais

Para o aumento dos acessos através das redes sociais, contribuiu o crescimento do Facebook e o aproveitamento desta rede social por parte do JPN para fazer a divulgação de notícias. Os primeiros acessos via Facebook datam de Setembro de 2008, com 3 sessões. Mas foi a partir de

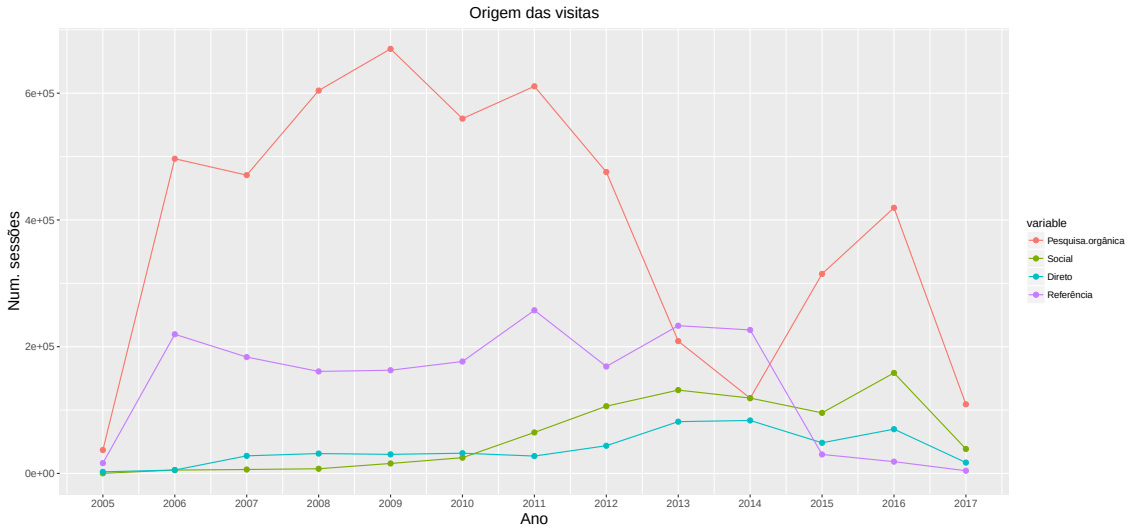

Figura 4.21: Distribuição do número de sessões em função do ano, tendo em conta a origem da visita

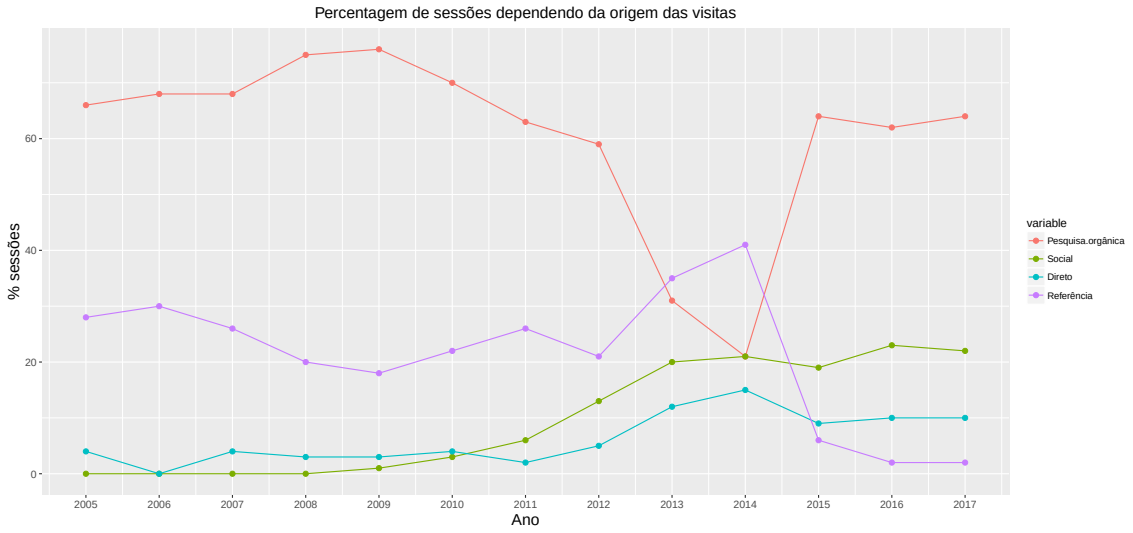

Figura 4.22: Distribuição da percentagem de sessões em função do ano, tendo em conta a origem da visita

Setembro de 2010 que o número de sessões com origem nessa rede social não baixou da ordem dos milhares, justificando assim o facto dos acessos via redes sociais terem ultrapassado os acessos diretos em 2011. No ano em as redes sociais passaram a ser as segundas maiores geradores de sessões no JPN, 2015, o Facebook deu origem a 94.782 sessões - 99,13% de todos as sessões por redes sociais e 19,6% do total de sessões desse ano. A Figura 4.23 mostra a evolução do número de sessões com origem no Facebook e a Figura 4.24 mostra a evolução da percentagem do total sessões que tiveram origem no Facebook.

A única rede social com acessos constantes de 2005 a 2017 é o Blogger, mas com uma notória

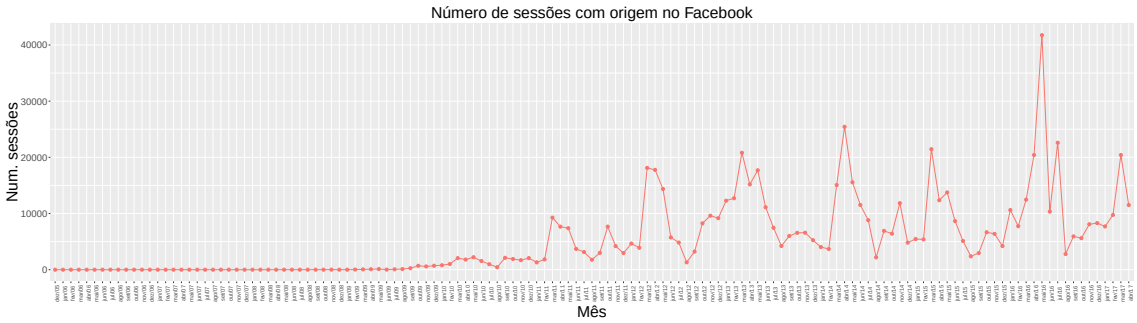

Figura 4.23: Número de sessões com origem no Facebook

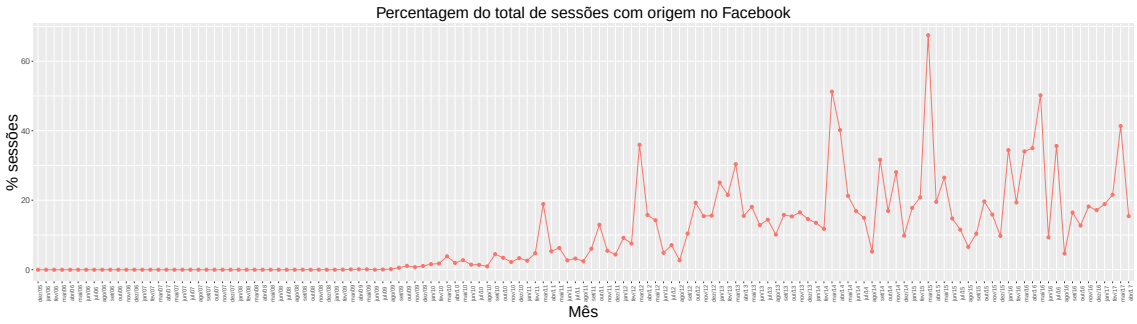

Figura 4.24: Percentagem do total de sessões com origem no Facebook

queda ao nível de acessos. No ano de 2006 deu origem a 3.434 sessões - 73,08% de todos as sessões por redes sociais e 0,46% do total de sessões desse ano. Já em 2016 deu origem a 233 sessões - 0,15% de todos as sessões por redes sociais e 0,035% do total de sessões desse ano. A Figura 4.25 mostra a evolução do número de sessões com origem no Blogger.

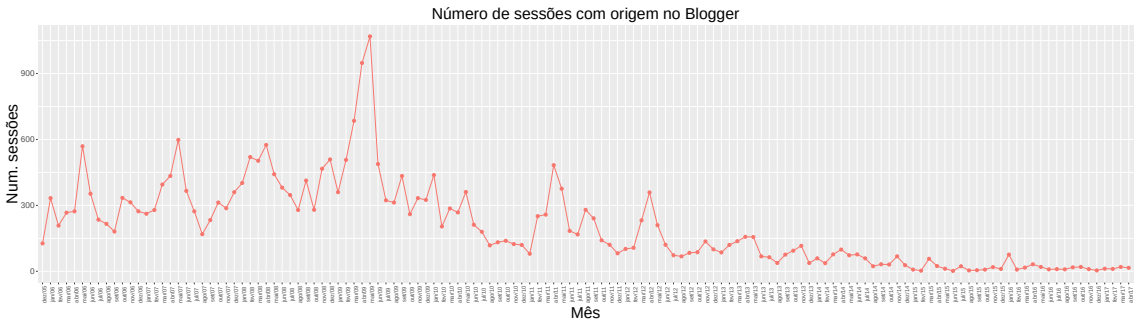

Figura 4.25: Número de sessões com origem no Blogger

Foram usadas um total de 56 redes sociais diferentes para aceder ao JPN, das quais as 10 mais usadas estão presentes na Figura 4.26. A Figura 4.27 oculta o Facebook e o Blogger para uma visão mais clara das restantes 8 redes sociais.

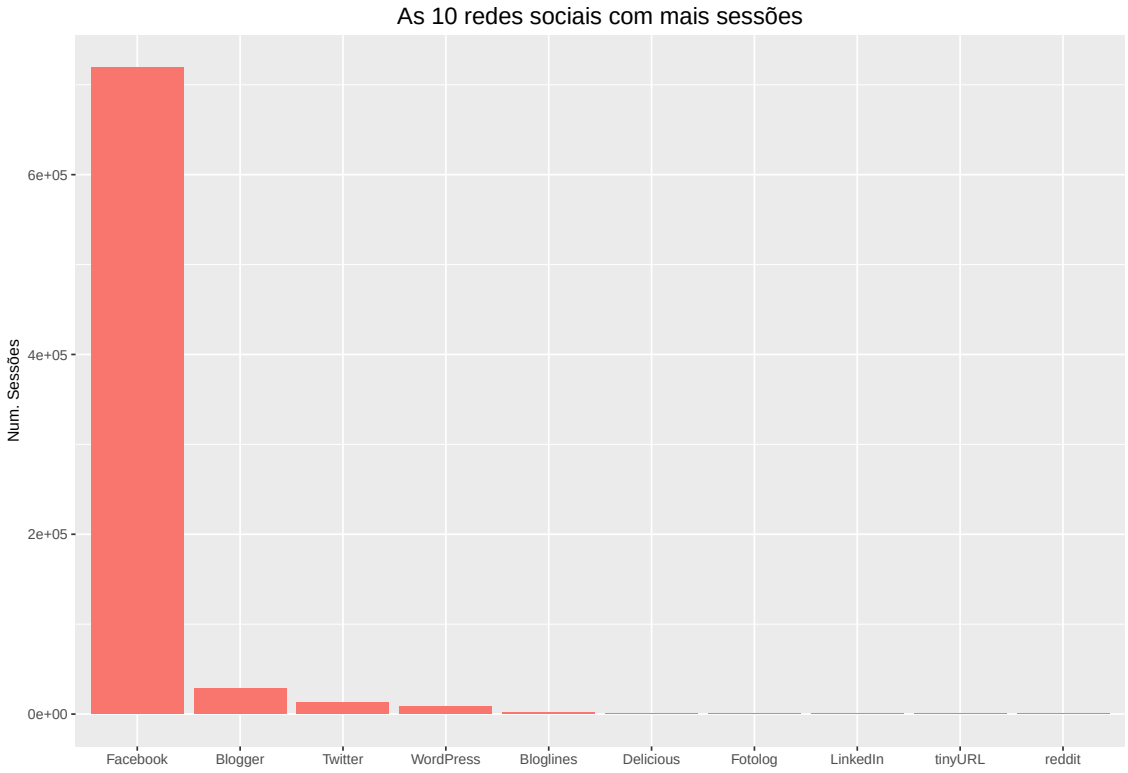

Figura 4.26: As 10 redes sociais com mais sessões

4.8 Momentos das visitas

Apesar das mudanças ocorridas ao nível dos dispositivos usados e da origem das visitas, as sessões em função do dia da semana mantiveram-se praticamente constantes ao longo deste grande período de tempo. A comparação feita é entre os anos de 2006, 2016 e o total de todos os anos. Há um ligeiro aumento na percentagem de sessões ocorridas entre segunda e quarta-feira, sendo que em 2016 esse aumento foi até quinta-feira. Sexta-feira, sábado e domingo são os dias com uma menor percentagem de sessões. A Figura 4.28 mostra a percentagem do total de sessões semanal em função do dia da semana. Através da análise das sessões em função do dia da semana, é mais uma vez visível o comportamento semelhante ao encontrado por Fang [Fan07] na sua análise de uma página com uma forte componente académica.

Também a percentagem de sessões em função da hora do dia tem tido um comportamento semelhante ao longo dos anos. Para o estudo desta métrica foram comparados os anos de 2007, 2016 e o total de todos os anos - o ano de 2006 ficou de fora, uma vez que a discrepância encontrada nos valores sugere que houve um problema de configuração nas horas do Google Analytics. As sessões começam a aumentar às 6h da manhã até perto do meio dia, altura em que ocorre uma quebra até às 13h, voltando a aumentar ligeiramente até às 15h. A partir daqui diminui até às 20h, voltando a aumentar entre as 21h e as 22h, altura em que, tanto no total como no ano de 2016, é atingida a maior percentagem de sessões. Após esta hora as sessões diminuem até às 6h. Olhando

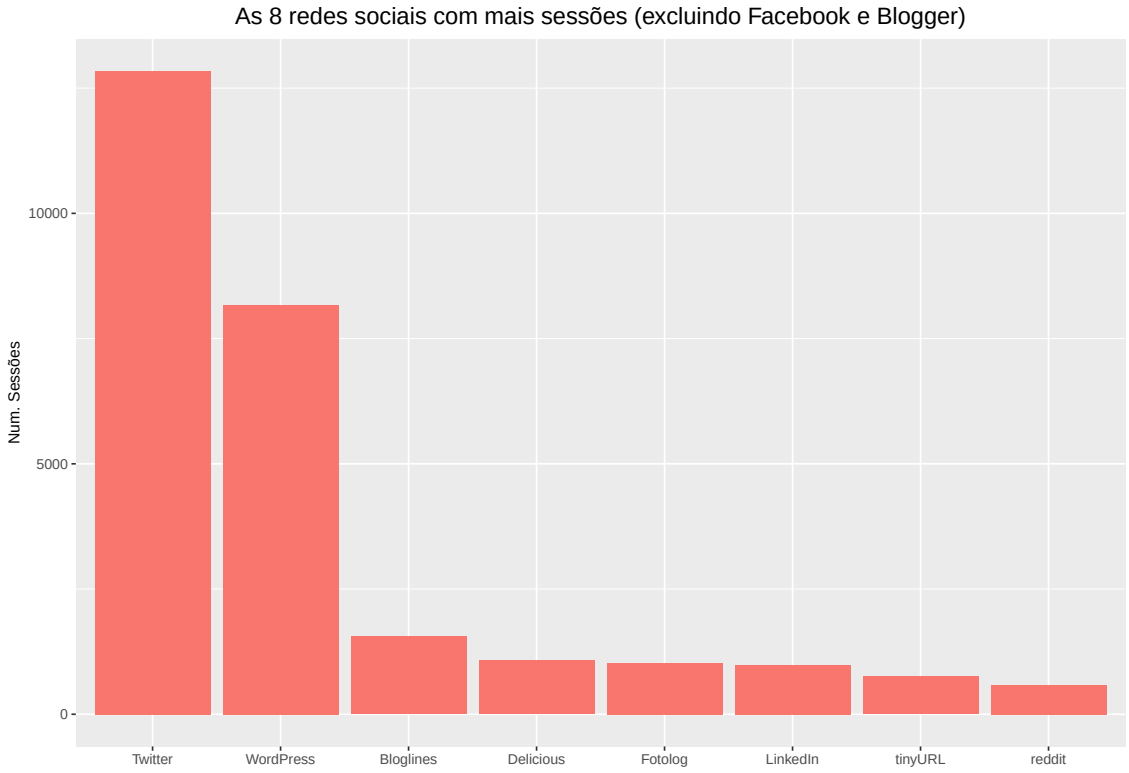

Figura 4.27: As 8 redes sociais com mais sessões (excluindo Facebook e Blogger)

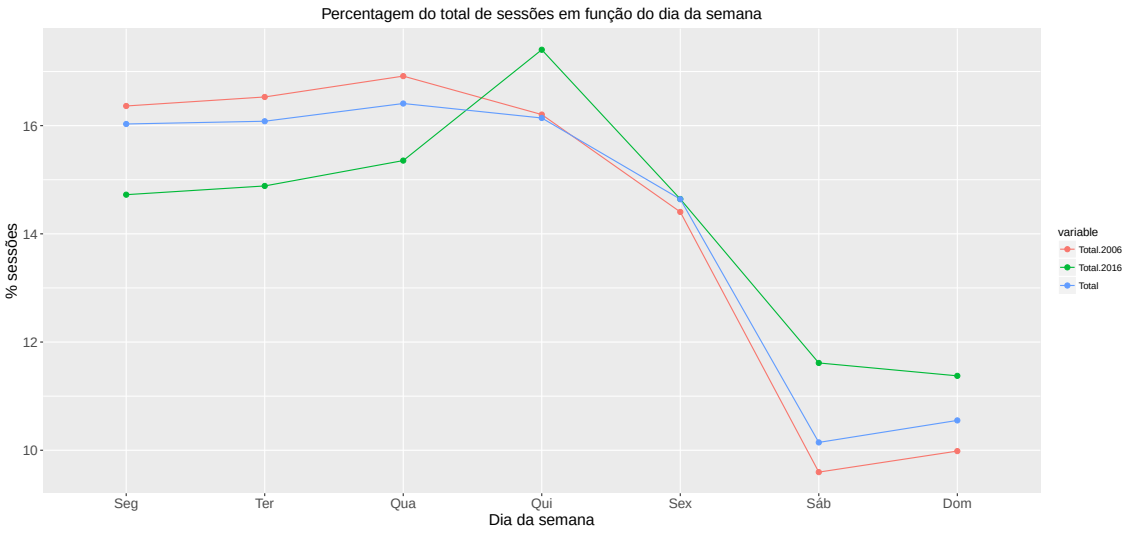

Figura 4.28: Percentagem do total de sessões em função do dia da semana

para o gráfico é facilmente identificável um dia de trabalho: o número de sessões aumenta durante a manhã, tendo a quebra na hora de almoço, e voltando a aumentar no período imediatamente após essa hora. À medida que se aproxima a hora de jantar o número de sessões volta a descer,

atingindo, posteriormente, valores mais elevados nas hora seguintes. Durante a noite a quebra é muito notória. A Figura 4.29 mostra a variação da percentagem do total de sessões diário em função da hora do dia.

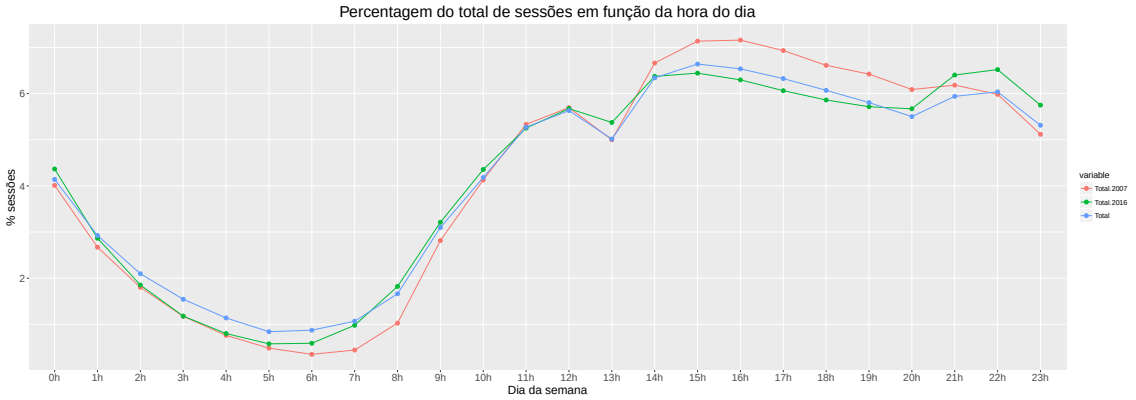

Figura 4.29: Percentagem do total de sessões em função da hora do dia

Especialmente interessante, embora as diferenças não sejam muito significativas e não tenha sido possível recolher estes dados de forma fidedigna, é a maneira como o maior uso de dispositivos móveis moldou os acessos às páginas do JPN. Grande parte das diferenças entre o ano de 2007 e 2016 encontram-se nas horas em que a maioria da população se dirige para as aulas ou empregos (das 7h às 9h) e no período após a hora de jantar, altura em que o uso dos dispositivos móveis se torna muito mais propício.

4.9 Pesquisa

A pesquisa só começou a ser monitorizada a partir de Abril de 2016. Neste período de um ano até Abril de 2017, houve apenas 3169 sessões com pesquisa, o que corresponde a 0,39% do total de sessões, com um total de 4563 pesquisas. O tempo médio da sessão após pesquisa (duração de todas as sessões com pesquisa/número de sessões com pesquisa) é muito superior ao da média do total de sessões - 4 minutos e 45 segundos contra 50 segundos. A Figura 4.30 mostra a diferença entre os tempos médios das sessões com e sem pesquisa.

Após uma pesquisa o número médio de páginas visitadas na sessão é de 2,86, também superior ao número médio total de páginas visitadas por sessão, 1,31. A Figura 4.31 mostra a diferença entre o número médio de páginas visitadas nas sessões com e sem pesquisa. 16,57% dos utilizadores saem da página de resultados de pesquisa sem visitar qualquer página apresentada. Quanto aos termos mais pesquisados, olhando para o top 5, é constituído por nomes de programas do JPN (PortOuvido, Gente Comum e Quatro em Linha) ou por termos usados em bastantes peças (noticiário e infografia). A Figura 4.32 mostra as 10 palavras mais pesquisadas no JPN. Apesar do JPN não ter um grande número de pesquisas, é possível concluir que os resultados apresentados são

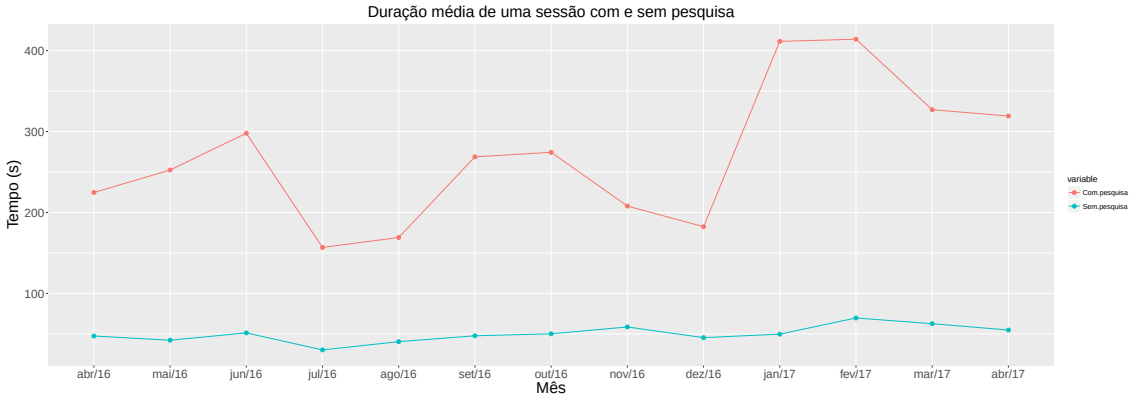

Figura 4.30: Duração média de uma sessão com e sem pesquisa

satisfatórios, uma vez que a taxa de saídas após uma pesquisa é relativamente baixa e a duração da sessão é quase 6 vezes superior a uma sessão sem pesquisa.

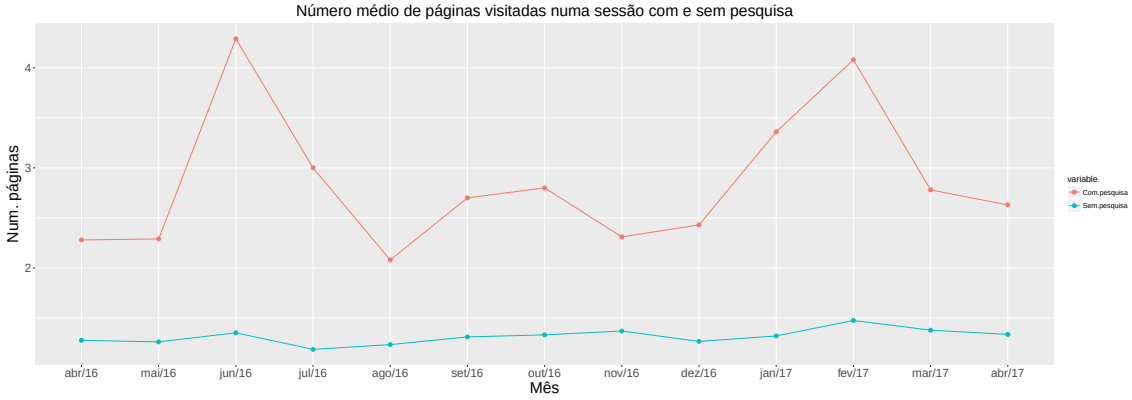

Figura 4.31: Número médio de páginas visitadas nas sessões com e sem pesquisa

4.10 Resultados dos inquéritos

Olhando para os resultados dos inquéritos, 100% das respostas foram dadas por estudantes, dos quais 77,8% estão na faixa etária entre os 18 e 24 anos e 22,2% entre os 25 e 34 anos. 77,8% estão no terceiro ano do curso e 22,2% no segundo. Relativamente ao interesse nas estatísticas das suas notícias 77,8% têm interesse em saber enquanto que 22,2% não têm interesse.

Na secção relativa às estatísticas de um artigo, os resultados encontram-se muito divididos, pelo que não se consegue perceber a relevância de algumas das métricas. O tempo médio de leitura é a métrica mais importante para os inquiridos, uma vez que 66,7% consideraram "Extremamente relevante", não tendo havido qualquer voto para "Nada relevante"ou "Pouco relevante". O número de visitas tem um total de 55,5% de votos acima de "Relevante"e 33,3% "Relevante". Já na idade do leitor e localização geográfica a média encontra-se no "Relevante". Finalmente, os dispositivos

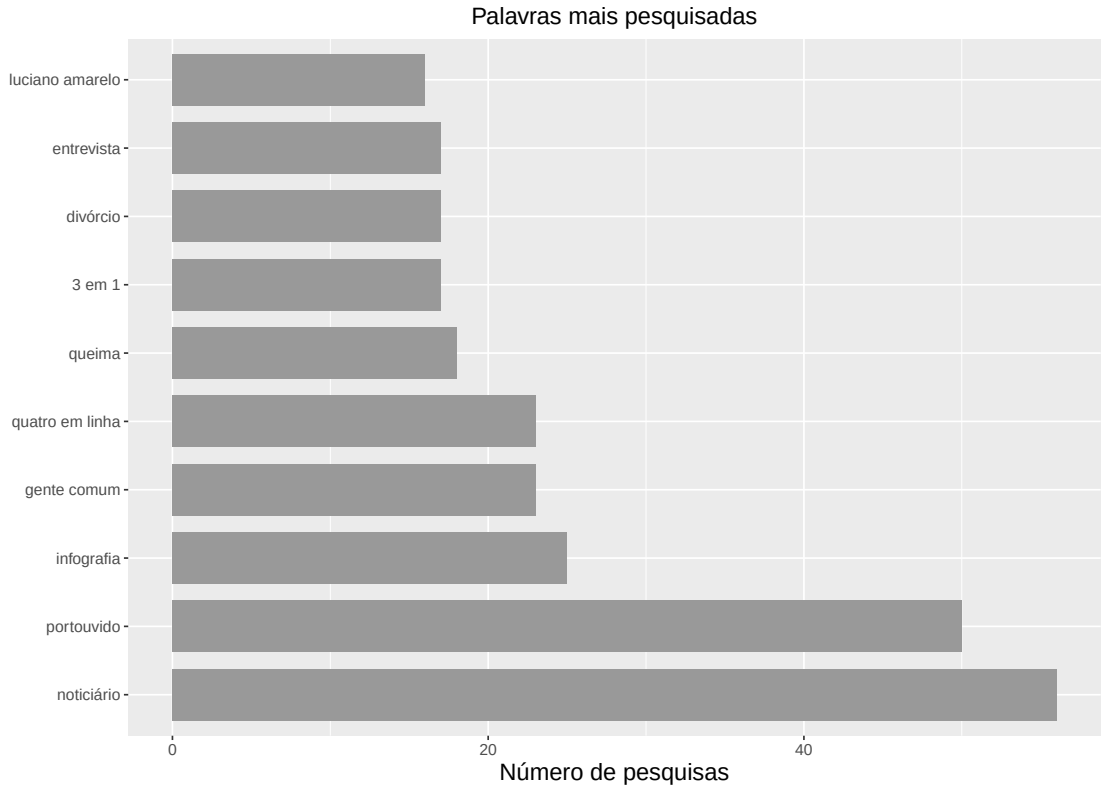

Figura 4.32: Top 10 das palavras pesquisadas

usados pelo leitor são a dimensão que, para os inquiridos, tem menos relevância uma vez que 33,3% consideram "Pouco relevante"e não há qualquer voto acima de "Relevante". Uma sugestão apresentada é a interação do leitor com o conteúdo (comentários, gostos, partilhas).

De seguida vem a secção relativa às estatísticas da página que apresenta as notícias de uma categoria. Nesta secção, a métrica mais relevante é a fidelização do leitor a uma categoria, com todos os votos iguais ou superiores a "Relevante", seguida do número de leitores que passaram a primeira página de resultados, com 88,8% dos votos iguais ou superiores a "Relevante". A idade do leitor, localização geográfica e o dispositivo usado têm uma média de respostas entre "Pouco relevante"e "Relevante".

Em relação à secção correspondente a todas as visitas ao jornal, o tempo de exploração da página de entrada, a fidelização do leitor à página e os termos mais usados em pesquisas no site são as métricas mais relevantes, com todos os votos iguais ou acima de "Relevante". O tempo de exploração da página de entrada teve 44,4% dos votos em "Extremamente relevante", 33,3% em "Muito relevante"e 22,2% em "Relevante". A fidelização do leitor teve 33,3% dos votos em "Extremamente relevante", 44,4% em "Muito relevante"e 22,2% em "Relevante". Os termos mais usados na pesquisa tiveram 22,2% dos votos em "Extremamente relevante"e em "Relevante"e 55,6% em "Muito relevante. A origem da visita tem 11,1% dos votos em "Nada relevante", sendo os restantes iguais ou acima de "Relevante". Já a idade, a localização e os dispositivos usados pelo

leitor têm uma média de "Relevante".

Apesar de algumas das métricas terem votos distribuídos pelas várias opções, com a média de votações a coincidirem com "Relevante", podemos concluir que, para os profissionais do meio jornalístico, as estatísticas relacionadas diretamente com o leitor não têm muita relevância - idade, localização, dispositivos usados, por exemplo. Uma métrica comum às secções, e também com resultados semelhantes, é o tempo despendido pelo utilizador numa determinada página, provavelmente de forma a perceber se o conteúdo está, de facto, a ser visualizado ou se a página apenas foi aberta e fechada de seguida. Também a fidelização dos leitores (medida através da percentagem de novas sessões) mostra ser importante, tanto a nível de categorias, como do conjunto de páginas do jornal. Os meios usados para a descoberta das página também têm grande relevância. É a partir da análise desta dimensão que é possível descobrir onde se pode aumentar a publicidade à página e onde se deve continuar o esforço já despendido. Finalmente, a análise dos termos das pesquisas também mostram uma relevância considerável. É a partir da análise destes termos que se sabe o que é que os utilizadores mais procuram. Os dados recolhidos no Google Analytics procuram satisfazer as necessidades encontradas através dos resultados dos inquéritos. No entanto, algumas das métricas não estão disponíveis para recolha, pelo que se tentou encontrar uma forma alternativa, por exemplo combinando dados de outras métricas, para alcançar os resultados esperados.

Capítulo 5

Conclusões

O ambiente académico em que o JPN se enquadra é bem visível ao analisar certos conjuntos de dados, como é o caso da oscilação das sessões de acordo com a época do ano e da idade dos visitantes. Também a fraca percentagem de retenção de utilizadores e, consequentemente, uma grande percentagem de novas sessões refletem esse ambiente, uma vez que, com o passar do tempo, os estagiários entram e saem do JPN. O mesmo acontece com os alunos da Universidade. A possibilidade de uma análise num intervalo temporal com uma dimensão tão grande permite uma previsão com uma maior percentagem de certezas sobre o que vai acontecer nos próximos anos. Se aquando da análise do tipo de dispositivos usados para aceder às páginas do JPN foi previsto que, a muito curto prazo, os dispositivos móveis iriam ser a principal fonte de acesso, no mês de Maio essa tendência veio mesmo a verificar-se, com 50,5% das sessões a terem origem em telemóveis e *tablets*.

Aproveitar a enorme expansão ocorrida no Facebook e fazer uso desta rede social para a divulgação das notícias revelou-se uma decisão acertada, uma vez que trouxe um aumento nas sessões com origem nas redes sociais. Perceber as tendências do mundo que nos rodeia tem uma grande importância para conseguir continuar a ser competitivo. No caso do aumento no uso de dispositivos móveis, há que procurar ver de que maneira se pode melhorar o que atualmente é publicado, e de que forma é publicado, ou até a organização das páginas. A orientação das fotografias que ilustram as notícias, por exemplo, é muito importante ao fazer a consulta num *desktop* ou em dispositivos móveis. Se a preferência fôr por fotografias na horizontal, há que considerar uma mudança para fotografias com orientação na vertical. Da mesma maneira há que pensar muito bem nos conteúdos à volta das notícias, uma vez que nos dispositivos móveis o espaço é mais reduzido.

Web analytics passa pela análise dos dados disponíveis e usar os resultados obtidos para melhorar a página web e, consequentemente, a experiência de utilização dos clientes, com o objetivo final de aumentar a rentabilidade da página. Conhecer os hábitos de utilização e de navegação dos utilizadores através do seu estudo é a principal forma de adaptar os conteúdos aos seus gostos e necessidades. É inegável que uma página corretamente adaptada aos seus utilizadores e com uma navegação intuitiva tem uma probabilidade muito mais alta de manter os seus utilizadores e atrair outros.

Nas reuniões efetuadas para avaliar quais as métricas e dimensões importantes, ou não, a recolher, um dos aspetos comuns a ser abordado foi a forma como o utilizador gere o tempo que passa a interagir com uma página. Os exemplos apresentados passam pela visualização de conteúdo multimédia, tanto vídeo como áudio, com principal incidência nas questões "Numa reportagem com várias pessoas entrevistadas, quantas foram ouvidas?", "Onde parou a visualização de determinado vídeo?", "O utilizador esteve, de facto, a ver o vídeo completo? Ou deixou-o apenas em reprodução?". Em relação a outro tipo de conteúdo, mais precisamente na página de entrada, o interesse principal passava por saber que percentagem da página tinha sido visualizada.

O Google Analytics apenas permitiu saber quanto tempo a página esteve em uso, não se obtendo os dados mais específicos. O maior registo de atividade é de 8 horas, 37 minutos e 45 segundos, sem a sessão expirar, mas não se sabe o que aconteceu. A notícia foi lida na íntegra? Esteve a ser escrito algum comentário? Serviu para um trabalho de pesquisa? A única coisa que se consegue saber é que nas cerca de 8h30 que a notícia esteve aberta houve interação com a página de, pelo menos, 29 minutos e 59 segundos em 29 minutos e 59 segundos.

Como trabalho futuro será interessante consultar utilizadores regulares do JPN e avaliar, inicialmente através de inquéritos, posteriormente através de contacto direto com quem der respostas relevantes para o estudo, de forma a perceber quais os seus hábitos de utilização e navegação em alguns dos conteúdos noticiosos apresentados.

Referências

[AIS93] Rakesh Agrawal, Tomasz Imieliński e Arun Swami. Mining association rules between sets of items in large databases. *Acm sigmod record, ACM*, 22(2):207–216, 1993.

[Apa] Apache. Log files - apache http server. [Online; acedido em Maio 16, 2017]. URL: https://httpd.apache.org/docs/1.3/logs.html#combined.

[AS+94] Rakesh Agrawal, Ramakrishnan Srikant et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[BP98] Sergey Brin e Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, Abril 1998.

[BRCA09] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha e Virgílio Almeida. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 49–62. ACM, 2009.

[BS02a] Paulo Batista e Mario J. Silva. Mining web access logs of an on-line newspaper. Departamento de Informática, Faculdade de Ciencias–Universidade de Lisboa. Lisboa. Portugal, 2002.

[BS02b] Paulo Batista e Mário J Silva. Mining web access logs of an on-line newspaper. *Recommendation and Personalization in eCommerce*, page 100, 2002.

[CD10] V Chitraa e Antony Selvadoss Davamani. An efficient path completion technique for web log mining. In *IEEE International Conference on Computational Intelligence and Computing Research*, 2010.

[Cha06] Li Chaofeng. Research and development of data preprocessing in web usage mining. In *International Conference on Management Science and Engineering*, 2006.

[CMS97] Robert Cooley, Bamshad Mobasher e Jaideep Srivastava. Web mining: Information and pattern discovery on the world wide web. In *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on*, pages 558–567. IEEE, 1997.

[Coo00] Robert Walker Cooley. *Web usage mining: discovery and application of interesting patterns from web data*. PhD thesis, Citeseer, 2000.

[DK10] Ms. Dipa Dixit e Ms. M Kiruthika. Preprocessing of web logs. *International Journal on Computer Science and Engineering*, 2(7):2447–2452, 2010.

[Etz96] Oren Etzioni. The world wide web: Quagmire or gold mine. *Communications of the ACM*, 39(11):65–68, Novembro 1996.

[Fan07] Wei Fang. Using google analytics for improving library website content and design: A case study. *J. Library Philosophy and Practice*, pages 1–17, 2007.

[Foua] The Apache Software Foundation. Apache poi - case studies. [Online; acedido em Junho 04, 2017]. URL: https://poi.apache.org/casestudies.html.

[Foub] The Apache Software Foundation. Apache poi - the java api for microsoft documents (mission statement). [Online; acedido em Junho 04, 2017]. URL: https://poi.apache.org/index.html#Mission+Statement.

[Fouc] The Apache Software Foundation. Poi-hssf and poi-xssf - java api to access microsoft excel format files. [Online; acedido em Junho 04, 2017]. URL: https://poi.apache.org/spreadsheet/index.html.

[FPSS96] Usama Fayyad, Gregory Piatetsky-Shapiro e Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[GMN11] L.K. Joshila Grace, V. Maheswari e Dhinaharan Nagamalai. Analysis of web logs and web user in web mining. *International Journal of Network Security & Its Applications*, 3(1):99–110, Janeiro 2011.

[Gooa] Google. Analytics help. [Online; acedido em Junho 05, 2017]. URL: https://support.google.com/analytics.

[Goob] Google. Comece a utilizar o google analytics. [Online; acedido em Junho 05, 2017]. URL: https://learndigital.withgoogle.com/atelierdigitalportugal/lesson/62.

[Hes12] Kirk M Hess. Discovering digital library user behavior with google analytics. *Code4Lib Journal*, 2012.

[HS95] Maurice Houtsma e Arun Swami. Set-oriented mining for association rules in relational databases. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, pages 25–33. IEEE, 1995.

[JK98] Anupam Joshi e Raghu Krishnapuram. Robust fuzzy clustering methods to support web mining. In *Proc. Workshop in Data Mining and knowledge Discovery, SIGMOD*, pages 15–23, 1998.

[JKA13] Mehak Jain, Mukesh Kumar e Naveen Aggarwal. Web usage mining: An analysis. *Journal of Emerging Technologies in Web Intelligence*, 5(3):240–246, Agosto 2013.

[JPNa] JPN. Jornalismo porto net. [Online; acedido em Maio 14, 2017]. URL: https://jpn.up.pt/.

[JPNb] JPN. Jornalismo porto net - estatuto editorial. [Online; acedido em Maio 14, 2017]. URL: https://jpn.up.pt/documentos/estatuto-editorial-do-jpn/.

[KB00] Raymond Kosala e Hendrik Blockeel. Web mining research: A survey. *SIGKDD Explorations Newsletter*, 2(1):1–15, Junho 2000.

[KND13] Ankit R Kharwar, Chandni A Naik e Niyanta K Desai. A complete pre processing method for web usage mining. *International Journal of Emerging Technology and Advanced Engineering*, pages 638–641, 2013.

[LJ12] Vijayashri Losarwar e Dr Madhuri Joshi. Data preprocessing in web usage mining. In *International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July*, pages 15–16, 2012.

[Mob05] Bamshad Mobasher. Web usage mining. In *Encyclopedia of Data Warehousing and Mining*, pages 1216–1220. IGI Global, 2005.

[Mob07] Bamshad Mobasher. Data mining for web personalization. In *The adaptive web*, pages 90–135, 2007.

[MTV95] Heikki Mannila, Hannu Toivonen e A Inkeri Verkamo. Discovering frequent episodes in sequences extended abstract. In *1st Conference on Knowledge Discovery and Data Mining*, 1995.

[MYTF02] Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi e Toshikazu Fukushima. Rmining product reputations on the web. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 341–349, 2002.

[OMS11] Mohammad Amin Omidvar, Vahid Reza Mirabi e Najes Shokry. Analyzing the impact of visitors on page views with google analytics. *arXiv preprint arXiv:1102.0735*, 2011.

[Pro] R Project. What is r? [Online; acedido em Junho 04, 2017]. URL: https://www.r-project.org/about.html.

[PS10] Ravi Kumar P e Ashutosh Kumar Singh. Web structure mining: Exploring hyperlinks and algorithms for information retrieval. *American Journal of applied sciences*, 7(6):840–845, 2010.

[Sar] Deepayan Sarkar. lattice: Trellis graphics for r. [Online; acedido em Junho 04, 2017]. URL: https://cran.r-project.org/web/packages/lattice/index.html.

[SCDT00] Jaideep Srivastava, Robert Cooley, Mukund Deshpande e Pang-Ning Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKKD Explorations*, 1(2):12–23, Janeiro 2000.

[SDK05] Jaideep Srivastava, Prasanna Desikan e Vipin Kumar. Web mining– concepts, applications and research directions. In *Foundations and Advances in Data Mining*, pages 279–312, 2005.

[SK97] Hidekazu Sakagami e Tomonari Kamba. Learning personal preferences on online newspaper articles from user behaviors. *Computer Networks and ISDN Systems*, 29(8):1447–1455, 1997.

[SZAS97] Cyrus Shahabi, Amir M Zarkesh, Jafar Adibi e Vishal Shah. Knowledge discovery from users web-page navigation. In *Research Issues in Data Engineering, 1997. Proceedings. Seventh International Workshop on*, pages 20–29. IEEE, 1997.

REFERÊNCIAS

[Tcw] R Core Team e contributors worldwide. The r base package. [Online; acedido em Junho 04, 2017]. URL: https://stat.ethz.ch/R-manual/R-devel/library/base/html/00Index.html.

[TR16] Y Thushara e V Ramesh. A study of web mining application on e-commerce using google analytics tool. *International Journal of Computer Applications*, 149(11):21–26, 2016.

[Wic10] Hadley Wickham. A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, 19(1):3–28, 2010.

[Wic13] Hadley Wickham. ggplot2, 2013. [Online; acedido em Junho 04, 2017]. URL: http://ggplot2.org/.

[WK09] Daniel Waisberg e Avinash Kaushik. Web analytics 2.0: empowering customer centricity. *The original Search Engine Marketing Journal*, 2(1):5–11, 2009.

Anexo A

Inquérito efetuado a pessoas da área do jornalismo

Análise comportamental de utilizadores de jornais online

Este questionário enquadra-se no âmbito da dissertação "Estudo e caraterização dos hábitos de utilização e navegação em jornais online". Os dados obtidos servem para avaliar a viabilidade do estudo de algumas métricas que permitem caraterizar um jornal online, através da análise estatística dos seus leitores, artigos e interação dos leitores com a página do jornal, neste trabalho aplicado ao JPN.

Após uma parte para caraterização do inquirido, o questionário encontra-se dividido em três seções, que correspondem a três níveis na organização da página web: artigos, categorias e página web. Em cada uma dessas seções são apresentadas métricas que devem ser classificadas, numa escala de 1 a 5, tendo em conta a sua relevância para caraterizar a seção onde está inserida.

Obrigado pela sua colaboração.

*Required

1. **Idade** *
   Mark only one oval.
   - ◯ 18-24
   - ◯ 25-34
   - ◯ 35-44
   - ◯ 45-54
   - ◯ 55+

2. **Ocupação** *
   Mark only one oval.
   - ◯ Estudante *Skip to question 5.*
   - ◯ Profissional *Skip to question 3.*

Profissional

3. **Anos de experiência** *
   Mark only one oval.
   - ◯ 4-
   - ◯ 5-10
   - ◯ 10-15
   - ◯ 15-20
   - ◯ 20+

4. **Gostaria de ter estatísticas sobre uma notícia sua?** *
   Mark only one oval.
   - ◯ Sim
   - ◯ Não

Skip to question 7.

Estudante

5. **Ano do curso** *
Mark only one oval.

- 1º
- 2º
- 3º

6. **Gostaria de ter estatísticas sobre uma notícia sua?** *
Mark only one oval.

- Sim
- Não

Artigos

Classifique as métricas apresentadas, de acordo com o que considera relevante para compreender o impacto do artigo junto dos leitores, tendo em conta a seguinte escala:

1 - Nada relevante
2 - Pouco relevante
3 - Relevante
4 - Muito relevante
5 - Extremamente relevante

7. **Número de visitas** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

8. **Idade do leitor** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

9. **Tempo médio de leitura** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

10. **Localização geográfica do leitor** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

11. **Dispositivo usado pelo leitor** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

12. **Sugestões**

Categorias

Classifique as métricas apresentadas, de acordo com o que considera relevante para compreender o impacto da página que apresenta os artigos de uma categoria, tendo em conta a seguinte escala:

1 - Nada relevante
2 - Pouco relevante
3 - Relevante
4 - Muito relevante
5 - Extremamente relevante

13. **Idade do leitor** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

14. **Localização geográfica do leitor** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

15. **Dispositivo usado pelo leitor** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

16. **Número de leitores que passaram a primeira página de resultados** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

17. **Fidelização do leitor a uma categoria (um leitor consulta a mesma categoria por diversas vezes)** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

18. **Sugestões**

Página Web

Classifique as seguintes métricas, de acordo com o que considera relevante, considerando todas as visitas ao jornal, tendo em conta a seguinte escala:

1 - Nada relevante
2 - Pouco relevante
3 - Relevante
4 - Muito relevante
5 - Extremamente relevante

19. **Idade do leitor** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

20. **Localização geográfica do leitor** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

21. **Dispositivo usado pelo leitor** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

22. **Tempo de exploração da homepage** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

23. **Fidelização do leitor à página (o leitor retorna à página do jornal)** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

24. **Origem da visita à página (redes sociais, pesquisa, entre outros)** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

25. **Termos mais usados em pesquisas na site** *
Mark only one oval.

1	2	3	4	5
◯	◯	◯	◯	◯

26. **Sugestões**

Fim

Obrigado pela sua colaboração!